A framework to characterize performance of LASSO algorithms

MIHAILO STOJNIC

School of Industrial Engineering
Purdue University, West Lafayette, IN 47907
e-mail: mstojnic@purdue.edu

Abstract

In this paper we consider solving *noisy* under-determined systems of linear equations with sparse solutions. A noiseless equivalent attracted enormous attention in recent years, above all, due to work of [12, 13, 25] where it was shown in a statistical and large dimensional context that a sparse unknown vector (of sparsity proportional to the length of the vector) can be recovered from an under-determined system via a simple polynomial ℓ_1-optimization algorithm. [13] further established that even when the equations are *noisy*, one can, through an SOCP noisy equivalent of ℓ_1, obtain an approximate solution that is (in an ℓ_2-norm sense) no further than a constant times the noise from the sparse unknown vector. In our recent works [62, 63], we created a powerful mechanism that helped us characterize exactly the performance of ℓ_1 optimization in the noiseless case (as shown in [61] and as it must be if the axioms of mathematics are well set, the results of [62, 63] are in an absolute agreement with the corresponding exact ones from [25]). In this paper we design a mechanism, as powerful as those from [62, 63], that can handle the analysis of a LASSO type of algorithm (and many others) that can be (or typically are) used for "solving" noisy under-determined systems. Using the mechanism we then, in a statistical context, compute the exact worst-case ℓ_2 norm distance between the unknown sparse vector and the approximate one obtained through such a LASSO. The obtained results match the corresponding exact ones obtained in [6, 26]. Moreover, as a by-product of our analysis framework we recognize existence of an SOCP type of algorithm that achieves the same performance.

Index Terms: **Noisy linear systems of equations; LASSO; SOCP; ℓ_1-optimization; compressed sensing**
.

1 Introduction

In recent years the problem of finding sparse solutions of under-determined systems of linear equations attracted enormous attention. Applications seem vast and as if they are growing almost on a daily basis (see, e.g. [4, 10, 14, 22, 30, 45, 49, 53, 55–57, 69, 71] and references therein). Given a substantial interest in the problem (and especially that it is coming from a variety of different fields), one may assume that designing efficient algorithms that would solve it could be of far-reaching importance. To that end, we believe that a precise mathematical understanding of the phenomena that make certain algorithms work well would help solidify belief in their success in current and future applications. Moreover, it is possible that down the road it can also help expand further the range of their applications.

Moving long the same lines, we in this paper focus on studying mathematical properties of under-determined systems of linear equations and certain algorithms used to solve them. We start the story by

introducing an idealized version of the problem that we plan to study. In its simplest form it amounts to finding a k-sparse $\mathbf{x}$ such that

$$A\mathbf{x} = \mathbf{y} \tag{1}$$

where A is an $m \times n$ ($m < n$) matrix and $\mathbf{y}$ is an $m \times 1$ vector (see Figure 1; here and in the rest of the paper, under k-sparse vector we assume a vector that has at most k nonzero components). Of course, the assumption will be that such an $\mathbf{x}$ exists (clearly, the case of real interest is $k < m$). To make writing in the rest of the paper easier, we will assume the so-called *linear* regime, i.e. we will assume that $k = \beta n$ and that the number of equations is $m = \alpha n$ where α and β are constants independent of n (more on the non-linear regime, i.e. on the regime when m is larger than linearly proportional to k can be found in e.g. [21, 34, 35]).
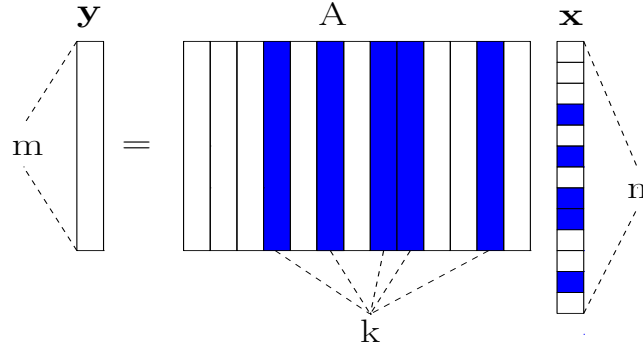

Figure 1: Model of a linear system; vector $\mathbf{x}$ is k-sparse

If one has the freedom to design matrix A then the results from [2, 46, 52] demonstrated that the techniques from coding theory (based on coding/decoding of Reed-Solomon codes) can be employed to determine *any* k-sparse $\mathbf{x}$ in (1) for any $0 < \alpha \leq 1$ and any $\beta \leq \frac{\alpha}{2}$ in polynomial time. It is relatively easy to show that under the unique recoverability assumption β can not be greater than $\frac{\alpha}{2}$. Therefore, as long as one is concerned with the unique recovery of k-sparse $\mathbf{x}$ in (1) in polynomial time the results from [2, 46, 52] are optimal. The complexity of algorithms from [2, 46, 52] is roughly $O(n^3)$. In a similar fashion one can, instead of using coding/decoding techniques associated with Reed/Solomon codes, design the matrix and the corresponding recovery algorithm based on the techniques related to coding/decoding of Expander codes (see e.g. [41, 42, 72] and references therein). In that case recovering $\mathbf{x}$ in (1) is significantly faster for large dimensions n. Namely, the complexity of the techniques from e.g. [41, 42, 72] (or their slight modifications) is usually $O(n)$ which is clearly for large n significantly smaller than $O(n^3)$. However, the techniques based on coding/decoding of Expander codes usually do not allow for β to be as large as $\frac{\alpha}{2}$.

On the other hand, if one has no freedom in choice of A designing the algorithms to find k-sparse $\mathbf{x}$ in (1) is substantially harder. In fact, when there is no choice in A the recovery problem (1) becomes NP-hard. Two algorithms 1) *Orthogonal matching pursuit - OMP* and 2) *Basis pursuit - ℓ_1-optimization* (and their different variations) have been often viewed historically as solid heuristics for solving (1) (in recent years belief propagation type of algorithms are emerging as strong alternatives as well). Roughly speaking, OMP algorithms are faster but can recover smaller sparsity whereas the BP ones are slower but recover higher sparsity. In a more precise way, under certain probabilistic assumptions on the elements of A it can be shown (see e.g. [51, 66, 67]) that if $m = O(k \log(n))$ OMP (or slightly modified OMP) can recover $\mathbf{x}$ in (1) with complexity of recovery $O(n^2)$. On the other hand a stage-wise OMP from [29] recovers $\mathbf{x}$ in (1) with complexity of recovery $O(n \log n)$. Somewhere in between OMP and BP are recent improvements CoSAMP (see e.g. [50]) and Subspace pursuit (see e.g. [23]), which guarantee (assuming the linear regime) that the k-sparse $\mathbf{x}$ in (1) can be recovered in polynomial time with $m = O(k)$ equations which is the same performance guarantee established in [13, 25] for the BP.

We now introduce the BP concept (or, as we will refer to it, the ℓ_1-optimization concept; a slight modification/adaptation of it will actually be the main topic of this paper). Variations of the standard ℓ_1-optimization from e.g. [15, 19, 60] as well as those from [24, 32, 37–39, 59] related to ℓ_q-optimization, $0 < q < 1$ are possible as well; moreover they can all be incorporated in what we will present below. The ℓ_1-optimization concept suggests that one can maybe find the k-sparse $\mathbf{x}$ in (1) by solving the following ℓ_1-norm minimization problem

$$
\begin{aligned}
\min \quad & \|\mathbf{x}\|_1 \\
\text{subject to} \quad & A\mathbf{x} = \mathbf{y}.
\end{aligned}
\tag{2}
$$

As is then shown in [13] if α and n are given, A is given and satisfies the restricted isometry property (RIP) (more on this property the interested reader can find in e.g. [1, 5, 11–13, 58]), then any unknown vector $\mathbf{x}$ with no more than $k = \beta n$ (where β is a constant dependent on α and explicitly calculated in [13]) non-zero elements can indeed be recovered by solving (2). In a statistical and large dimensional context in [25] and later in [63] for any given value of β the exact value of the maximum possible α was determined.

As we mentioned earlier the above scenario is in a sense idealistic. Namely, it assumes that $\mathbf{y}$ in (2) was obtained through (1). On the other hand in many applications only a *noisy* version of $A\mathbf{x}$ may be available for $\mathbf{y}$ (this is especially so in measuring type of applications) see, e.g. [12, 13, 40, 70]. When that happens one has the following equivalent to (1) (see, Figure 2)

$$
\mathbf{y} = A\mathbf{x} + \mathbf{v},
\tag{3}
$$

where $\mathbf{v}$ is an $m \times 1$ vector (often dubbed as the noise vector; the so-called ideal case presented above is of course a special case of the noisy case). Finding the k-sparse $\mathbf{x}$ in (3) is now incredibly hard. Basically, one
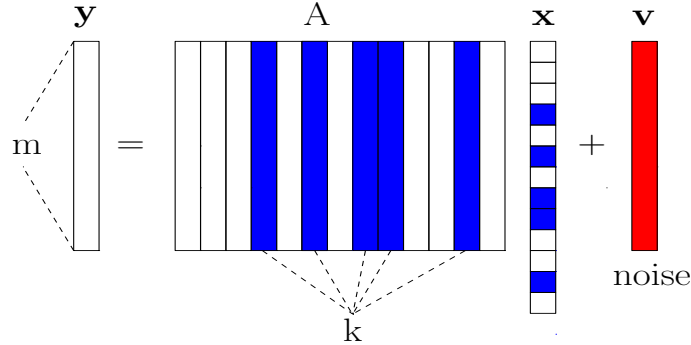

Figure 2: Model of a linear system; vector $\mathbf{x}$ is k-sparse

is looking for a k-sparse $\mathbf{x}$ such that (3) holds and on top of that $\mathbf{v}$ is unknown. Although the problem is hard there are various heuristics throughout the literature that one can use to solve it approximately. Below we restrict our attention to two groups of algorithms that we believe are the most relevant to the results that we will present.

To introduce a bit or tractability in finding the k-sparse $\mathbf{x}$ in (3) one usually assumes certain amount of knowledge about either $\mathbf{x}$ or $\mathbf{v}$. As far as tractability assumptions on $\mathbf{v}$ are concerned one typically (and possibly fairly reasonably in applications of interest) assumes that $\|\mathbf{v}\|_2$ is bounded (or highly likely to be bounded) from above by a certain known quantity. The following second-order cone programming (SOCP)

analogue to (2) is one of the approaches that utilizes such an assumption (see, e.g. [13])

$$\min_{\mathbf{x}} \quad \|\mathbf{x}\|_1$$
$$\text{subject to} \quad \|\mathbf{y} - A\mathbf{x}\|_2 \le r \tag{4}$$

where, r is a quantity such that $\|\mathbf{v}\|_2 \le r$ (or r is a quantity such that $\|\mathbf{v}\|_2 \le r$ is say highly likely). For example, in [13] a statistical context is assumed and based on the statistics of $\mathbf{v}$, r was chosen such that $\|\mathbf{v}\|_2 \le r$ happens with overwhelming probability (as usual, under overwhelming probability we in this paper assume a probability that is no more than a number exponentially decaying in n away from 1). Given that (4) is now among few almost standard choices when it comes to finding the $\mathbf{x}$-sparse in (3), the literature on its properties is vast (see, e.g. [13, 28, 65] and references therein). Also, given that this SOCP will not be the main topic of this paper we below briefly mention only what we consider to be the most influential work on this topic in recent years. Namely, in [13] the authors analyzed performance of (4) and showed a result similar in flavor to the one that holds in the ideal - noiseless - case. In a nutshell the following was shown in [13]: let $\mathbf{x}$ be a βn-sparse vector such that (3) holds and let $\mathbf{x}_{socp}$ be the solution of (4). Then $\|\mathbf{x}_{socp} - \mathbf{x}\|_2 \le Cr$ where β is a constant independent of n and C is a constant independent of n and of course dependent on α and β. This result in a sense establishes a noisy equivalent to the fact that a linear sparsity can be recovered from an under-determined system of linear equations. In an informal language, it states that a linear sparsity can be *approximately* recovered in polynomial time from a noisy under-determined system with the norm of the recovery error guaranteed to be within a constant multiple of the noise norm. Establishing such a result is, of course, a feat in its own class, not only because of its technical contribution but even more so because of the amount of interest that it generated in the field.

In this paper we will also consider an approximate recovery of the k-sparse $\mathbf{x}$ in (3). However, instead of the above mentioned SOCP we will focus on a group of highly successful algorithms called LASSO (the LASSO algorithms, as well as the SOCP ones, are of course well known in the statistics community and there is again a vast literature that covers their performance (see, e.g. [6,9,17,18,26,48,64,68] and references therein). There are many variants of LASSO but the following one is probably the most well known

$$\min_{\mathbf{x}} \|\mathbf{y} - A\mathbf{x}\|_2^2 + \lambda_{lasso}\|\mathbf{x}\|_1. \tag{5}$$

λ_{lasso} in (5) is a parameter to be chosen based on the amount of pre-knowledge one may have about A, $\mathbf{v}$, and/or $\mathbf{x}$. The results that relate to the characterization of the approximation error of (5) that are similar to the SOCP ones mentioned above can be established (see, e.g. [7]). Of course, characterizing the performance of the recovery algorithm through the norm-2 of the error vector is only one possible way among many (more on other measures of performance can be found in e.g. [9, 70]). In this paper we will develop a novel framework for performance characterization of the LASSO algorithms. Among other things, in a statistical context, the framework will enable us to provide a precise characterization of the norm-2 of the approximation error of the LASSO algorithms.

While our main focus in this paper are algorithms from the LASSO group we mention that besides the SOCP and LASSO algorithms there are of course various other algorithms/heuristics that have been suggested as possible alternatives throughout the literature in recent years. Such an alternative that gained certain amount of popularity is for example the so-called Dantzig selector introduced in [16]. The Dantzig selector amounts to solving the following optimization problem

$$\min \quad \|\mathbf{x}\|_1$$
$$\text{subject to} \quad \|A^T(A\mathbf{x} - \mathbf{y})\|_\infty \le C_{Dan},$$

where C_{Dan} is a carefully chosen parameter that of course should depend on A, $\mathbf{v}$, and/or $\mathbf{x}$. As a lin-

ear program the Danzig selector promises to be faster than SOCP or LASSO which are both quadratic programs. On the other hand recent improvements in numerical implementations of LASSO's and their solid approximate recovery abilities make them quite competitive as well (more on a thorough discussion/comparison, advantages/disdvatnages of the Dantzig selector and the LASSO algorithms can be found in e.g. [3, 8, 31, 33, 43, 44, 47]).

To facilitate the exposition and the easiness of following we will present our framework on a version of the LASSO from (5). Namely, we will consider,

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & \|\mathbf{y} - A\mathbf{x}\|_2 \\
\text{subject to} \quad & \|\mathbf{x}\|_1 \leq \|\tilde{\mathbf{x}}\|_1
\end{aligned} \tag{6}
$$

where $\tilde{\mathbf{x}}$ is the original k-sparse $\mathbf{x}$ that satisfies (3) (we just briefly mention that in a context that will be considered in this paper it is not that difficult to transform the LASSO from (6) to one that is structurally equivalent to (5); however, we stop short of exploring this connection further before presenting our main results and only mention that a section towards the end of the paper will explore it in more detail.). We do however mention right here that in order to run (6) one does require the knowledge of $\|\tilde{\mathbf{x}}\|_1$. In a sense this requirement is an equivalent to setting r and λ_{lasso} in (4) and (5), respectively. In order to be maximally effective both r and λ_{lasso} do require some amount of pre-knowledge about A, $\mathbf{v}$, and/or $\mathbf{x}$.

Before we proceed further we briefly summarize the organization of the rest of the paper. In Section 2, we present a statistical framework for the performance analysis of the LASSO algorithms. To demonstrate its power we towards the end of Section 2, for any given α and β, compute the worst case norm-2 of the error that (6) makes when used for approximate recovery of general sparse signals $\mathbf{x}$ from (3). In Section 3 we then specialize results from Section 2 to the so-called signed vectors $\mathbf{x}$. In Section 4 we discuss how the LASSO from (6) can be connected to the LASSO from (5). In Section 5 we demonstrate that there is an SOCP algorithm (similar to the one given in (4)) that achieves the same performance as do (6) and a corresponding (5). In Section 6 we present results that we obtained through numerical experiments. Finally, in Section 7 we discuss obtained results.

2 LASSO's performance analysis framework – general $\mathbf{x}$

In this section we create a statistical LASSO's performance analysis framework. Before proceeding further we will now explicitly state the major assumptions that we will make (the remaining ones, will be made appropriately throughout the analysis). Namely, in the rest of the paper we will assume that the elements of A are i.i.d. standard normal random variables. We will also assume that the elements of $\mathbf{v}$ are i.i.d. Gaussian random variables with zero mean and variance σ. As stated earlier, we will assume that $\tilde{\mathbf{x}}$ is the original $\mathbf{x}$ in (3) that we are trying to recover and that it is *any* k-sparse vector with a given fixed location of its nonzero elements and a given fixed combination of their signs. Since the analysis (and the performance of (6)) will clearly be irrelevant with respect to what particular location and what particular combination of signs of nonzero elements are chosen, we can for the simplicity of the exposition and without loss of generality assume that the components $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_{n-k}$ of $\mathbf{x}$ are equal to zero and the components $\mathbf{x}_{n-k+1}, \mathbf{x}_{n-k+2}, \dots, \mathbf{x}_n$ of $\mathbf{x}$ are greater than or equal to zero. Moreover, throughout the paper we will call such an $\mathbf{x}$ k-sparse and positive. In a more formal way we will set

$$
\begin{aligned}
& \tilde{\mathbf{x}}_1 = \tilde{\mathbf{x}}_2 = \cdots = \tilde{\mathbf{x}}_{n-k} = 0 \\
& \tilde{\mathbf{x}}_{n-k+1} \geq 0, \tilde{\mathbf{x}}_{n-k+1} \geq 0, \dots, \tilde{\mathbf{x}}_n \geq 0.
\end{aligned} \tag{7}
$$

We also now take the opportunity to point out a rather obvious detail. Namely, the fact that $\tilde{\mathbf{x}}$ is positive is assumed for the purpose of the analysis. However, this fact is not known *a priori* and is not available to the solving algorithm (this will of course change in Section 3).

Once we establish the framework it will be clear that it can be used to characterize many of the LASSO features. We will defer these details to a collection of forthcoming papers. In this paper we will present only a small application that relates to a classical question of quantifying the approximation error that (6) makes when used to recover *any* k-sparse $\mathbf{x}$ that satisfies (3) and is from a set of $\mathbf{x}$'s with a given fixed location of nonzero elements and a given fixed combination of their signs.

Before proceeding further we will introduce a few definitions that will be useful in formalizing this application as well as in conducting the entire analysis. As it is natural we start with the solution of (6). Let $\hat{\mathbf{x}}$ be the solution of (6) and let $\mathbf{w}_{lasso} \in R^n$ be such that

$$\hat{\mathbf{x}} = \tilde{\mathbf{x}} + \mathbf{w}_{lasso}. \tag{8}$$

As an application of our framework we will compute the largest possible value of $\|\hat{\mathbf{x}} - \tilde{\mathbf{x}}\|_2 = \|\mathbf{w}_{lasso}\|_2$ for any combination (α, β). Or more rigorously, for any combination (α, β), we will find a d_{lasso} such that

$$\lim_{n \to \infty} P(d_{lasso} - \epsilon \leq \max_{\tilde{\mathbf{x}}} \|\mathbf{w}_{lasso}\|_2 \leq d_{lasso} + \epsilon) = 1 \tag{9}$$

for an arbitrarily small constant ϵ. However, before doing so we will first present the general framework. The framework that we will present will center around finding the optimal value of the objective function in (6) (of course in a probabilistic context). In the first of the following two subsections we will create a lower bound on this optimal value. We will then afterwards in the second of the subsections create an upper bound on this optimal value. Naturally in the third subsection we will show that the two bounds actually match. To make further writing easier and clearer we set already here

$$\begin{aligned} \zeta_{obj} = \min_{\mathbf{x}} \quad & \|\mathbf{y} - A\mathbf{x}\|_2 \\ \text{subject to} \quad & \|\mathbf{x}\|_1 \leq \|\tilde{\mathbf{x}}\|_1. \end{aligned} \tag{10}$$

2.1 Lower-bounding ζ_{obj}

In this section we present the part of the framework that relates to finding a "high-probability" lower bound on ζ_{obj}. To make arguments that will follow less tedious we will make an assumption that is significantly weaker than what we will eventually prove. Namely, we will assume that there is a (if necessary, arbitrarily large) constant $C_{\mathbf{w}}$ such that

$$P(\|\mathbf{w}_{lasso}\|_2 \leq C_{\mathbf{w}}) \geq 1 - e^{-\epsilon C_{\mathbf{w}} n}. \tag{11}$$

To make our arguments flow more naturally, one should probably provide a direct proof of this statement right here. However, given the difficulty of the task ahead we refrain from that and assume that the statement is correct. Roughly speaking, what we assume is that $\|\mathbf{w}_{lasso}\|_2$ is bounded by an arbitrarily large constant (of course we hope to create a machinery that can prove much more than (11)).

We start by noting that if one knows that $\mathbf{y} = A\tilde{\mathbf{x}} + \mathbf{v}$ holds then (10) can be rewritten as

$$\begin{aligned} \min_{\mathbf{x}} \quad & \|\mathbf{v} + A\tilde{\mathbf{x}} - A\mathbf{x}\|_2 \\ \text{subject to} \quad & \|\mathbf{x}\|_1 \leq \|\tilde{\mathbf{x}}\|_1. \end{aligned} \tag{12}$$

After a small change of variables, $\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{w}$, (12) becomes

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|\mathbf{v} - A\mathbf{w}\|_2 \\ \text{subject to} \quad & \|\tilde{\mathbf{x}} + \mathbf{w}\|_1 \leq \|\tilde{\mathbf{x}}\|_1, \end{aligned} \qquad (13)$$

or in a more compact form

$$\begin{aligned} \min_{\mathbf{w}} \quad & \|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \|_2 \\ \text{subject to} \quad & \|\tilde{\mathbf{x}} + \mathbf{w}\|_1 \leq \|\tilde{\mathbf{x}}\|_1, \end{aligned} \qquad (14)$$

where $A_{\mathbf{v}} = \begin{bmatrix} -A & \mathbf{v} \end{bmatrix}$ is now an $m \times (n+1)$ random matrix with i.i.d. standard normal components. Let

$$S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}}) = \{ \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \in R^{n+1} | \quad \|\mathbf{w}\|_2 \leq C_{\mathbf{w}} \quad \text{and} \quad \|\tilde{\mathbf{x}} + \mathbf{w}\|_1 \leq \|\tilde{\mathbf{x}}\|_1 \}. \qquad (15)$$

Further, let

$$f_{obj}(\sigma, \mathbf{w}) = \|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \|_2 \qquad (16)$$

and set,

$$\zeta_{obj}^{(help)} = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} f_{obj}(\sigma, \mathbf{w}) = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \|_2 = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \max_{\|\mathbf{a}\|_2=1} \mathbf{a}^T A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix}.$$
$$(17)$$

We now state a lemma from [36] that will be of use in what follows.

Lemma 1. *([36]) Let A be an $m \times n$ matrix with i.i.d. standard normal components. Let $\mathbf{g}$ and $\mathbf{h}$ be $m \times 1$ and $n \times 1$ vectors, respectively, with i.i.d. standard normal components. Also, let g be a standard normal random variable and let $\Phi \subset R^n$ be an arbitrary subset. Then for all choices of real ψ_ϕ*

$$P(\min_{\phi \in \Phi} \max_{\|\mathbf{a}\|_2=1} (\mathbf{a}^T A \phi + \|\phi\|_2 g - \psi_\phi) \geq 0) \geq P(\min_{\phi \in \Phi} \max_{\|\mathbf{a}\|_2=1} (\|\phi\|_2 \sum_{i=1}^{m} \mathbf{g}_i \mathbf{a}_i + \sum_{i=1}^{n} \mathbf{h}_i \phi_i - \psi_\phi) \geq 0). \quad (18)$$

Now, after applying Lemma 1 one has

$$P(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} (f_{obj}(\sigma, \mathbf{w}) + \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} g) \geq \zeta_{obj}^{(l)})$$

$$= P\left(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \max_{\|\mathbf{a}\|_2=1} \left(\mathbf{a}^T A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} + \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} g \right) \geq \zeta_{obj}^{(l)} \right)$$

$$\geq P\left(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \max_{\|\mathbf{a}\|_2=1} \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \sum_{i=1}^{m} \mathbf{g}_i \mathbf{a}_i + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i + \mathbf{h}_{n+1} \sigma \right) \geq \zeta_{obj}^{(l)} \right). \quad (19)$$

In what follows we will analyze the following probability

$$p_l = P\left(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \max_{\|\mathbf{a}\|_2=1} \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \sum_{i=1}^{m} \mathbf{g}_i \mathbf{a}_i + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i + \mathbf{h}_{n+1} \sigma \right) \geq \zeta_{obj}^{(l)} \right), \qquad (20)$$

which is of course nothing but the probability on the left-hand side of the inequality in (19). We will

7

essentially show that for certain $\zeta_{obj}^{(l)}$ this probability is close to 1. That will rather obviously imply that we have a "high probability" lower bound on ζ_{obj}. To that end, we first note that the maximization over $\mathbf{a}$ is trivial and one obtains

$$p_l = P\left(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})}\left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n}\mathbf{h}_i\mathbf{w}_i\right) + \mathbf{h}_{n+1}\sigma \geq \zeta_{obj}^{(l)}\right). \tag{21}$$

To facilitate the exposition that will follow let

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})}\left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n}\mathbf{h}_i\mathbf{w}_i\right). \tag{22}$$

One should note here that, although present in the definition of $S_{\mathbf{w}}$, σ clearly does not have an impact on the result of the above optimization. Now we split the analysis into two parts. The first one will be the deterministic analysis of $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ and will be presented in Subsection 2.1.1. In the second part (that will be presented in Subsection 2.1.2) we will use the results of such a deterministic analysis and continue the above probabilistic analysis applying various concentration results.

2.1.1   Optimizing $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$

In this section we compute $\xi(\sigma, \mathbf{g}, \mathbf{h})$. We first rewrite the optimization problem from (22) in the following possibly clearer form

$$
\begin{aligned}
\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{\mathbf{w}} \quad & \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n}\mathbf{h}_i\mathbf{w}_i \\
\text{subject to} \quad & \|\tilde{\mathbf{x}} + \mathbf{w}\|_1 \leq \|\tilde{\mathbf{x}}\|_1 \\
& \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \leq \sqrt{C_{\mathbf{w}}^2 + \sigma^2}.
\end{aligned}
\tag{23}
$$

To remove the absolute values we introduce auxiliary variables $\mathbf{t}_i, 1 \leq i \leq n$ and transform the above problem to

$$
\begin{aligned}
\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{\mathbf{w}, \mathbf{t}} \quad & \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n}\mathbf{h}_i\mathbf{w}_i \\
\text{subject to} \quad & \sum_{i=1}^{n}\mathbf{t}_i \leq \|\tilde{\mathbf{x}}\|_1 \\
& \tilde{\mathbf{x}}_i + \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n \\
& -\tilde{\mathbf{x}}_i - \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n \\
& \mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k \\
& -\mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k \\
& \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \leq \sqrt{C_{\mathbf{w}}^2 + \sigma^2}.
\end{aligned}
\tag{24}
$$

8

The Lagrange dual of the above problem then becomes

$$\mathcal{L}(\nu, \lambda^{(1)}, \lambda^{(2)}, \mathbf{w}, \mathbf{t}, \gamma) = \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n}\mathbf{h}_i\mathbf{w}_i + \nu\sum_{i=1}^{n}\mathbf{t}_i - \nu\|\tilde{\mathbf{x}}\|_1 + \sum_{i=n-k+1}^{n}\lambda_i^{(1)}(\tilde{\mathbf{x}}_i + \mathbf{w}_i - \mathbf{t}_i)$$

$$+ \sum_{i=n-k+1}^{n}\lambda_i^{(2)}(-\tilde{\mathbf{x}}_i - \mathbf{w}_i - \mathbf{t}_i) + \sum_{i=1}^{n-k}\lambda_i^{(1)}(\mathbf{w}_i - \mathbf{t}_i) + \sum_{i=1}^{n-k}\lambda_i^{(2)}(-\mathbf{w}_i - \mathbf{t}_i) + \gamma(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} - \sqrt{C_{\mathbf{w}}^2 + \sigma^2}). \tag{25}$$

After rearranging the terms we further have

$$\mathcal{L}(\nu, \lambda^{(1)}, \lambda^{(2)}, \mathbf{w}, \mathbf{t}, \gamma) = \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n}\mathbf{h}_i\mathbf{w}_i - \nu\|\tilde{\mathbf{x}}\|_1 + \sum_{i=1}^{n}\mathbf{t}_i(\nu - \lambda_i^{(1)} - \lambda_i^{(2)}) + \sum_{i=n-k+1}^{n}\lambda_i^{(1)}(\tilde{\mathbf{x}}_i + \mathbf{w}_i)$$

$$+ \sum_{i=n-k+1}^{n}\lambda_i^{(2)}(-\tilde{\mathbf{x}}_i - \mathbf{w}_i) + \sum_{i=1}^{n-k}\lambda_i^{(1)}\mathbf{w}_i - \sum_{i=1}^{n-k}\lambda_i^{(2)}\mathbf{w}_i + \gamma(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} - \sqrt{C_{\mathbf{w}}^2 + \sigma^2}). \tag{26}$$

After a few further arrangements we finally have

$$\mathcal{L}(\nu, \lambda^{(1)}, \lambda^{(2)}, \mathbf{w}, \mathbf{t}, \gamma) = \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}(\|\mathbf{g}\|_2 + \gamma) + \sum_{i=1}^{n}\mathbf{h}_i\mathbf{w}_i - \nu\|\tilde{\mathbf{x}}\|_1 + \sum_{i=1}^{n}\mathbf{t}_i(\nu - \lambda_i^{(1)} - \lambda_i^{(2)})$$

$$+ \sum_{i=n-k+1}^{n}(\lambda_i^{(1)} - \lambda_i^{(2)})\tilde{\mathbf{x}}_i + \sum_{i=1}^{n}(\lambda_i^{(1)} - \lambda_i^{(2)})\mathbf{w}_i - \gamma\sqrt{C_{\mathbf{w}}^2 + \sigma^2}. \tag{27}$$

Setting $(\nu - \lambda_i^{(1)} - \lambda_i^{(2)}) = 0, 1 \leq i \leq n$, (to insure that the dual is bounded) and combining (24) and (27) is enough to obtain

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(1)}, \lambda^{(2)}, \gamma} \min_{\mathbf{w}, \mathbf{t}} \quad \mathcal{L}(\nu, \lambda^{(1)}, \lambda^{(2)}, \mathbf{w}, \mathbf{t})$$

$$\text{subject to} \quad \lambda_j^{(i)} \geq 0, 1 \leq j \leq n, 1 \leq i \leq 2$$
$$\nu \geq 0$$
$$\nu - \lambda_i^{(1)} - \lambda_i^{(2)} = 0, 1 \leq i \leq n$$
$$\gamma \geq 0, \tag{28}$$

where we of course use the fact that the strict duality obviously holds. After removing the minimization over $\mathbf{t}$ we have

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(1)}, \lambda^{(2)}, \gamma} \min_{\mathbf{w}} \quad \mathcal{L}(\nu, \lambda^{(1)}, \lambda^{(2)}, \mathbf{w}, \gamma)$$

$$\text{subject to} \quad \lambda^{(i)} \geq 0, 1 \leq j \leq n, 1 \leq i \leq 2$$
$$\nu \geq 0$$
$$\nu - \lambda_i^{(1)} - \lambda_i^{(2)} = 0, 1 \leq i \leq n$$
$$\gamma \geq 0. \tag{29}$$

9

where

$$\mathcal{L}(\nu, \lambda^{(1)}, \lambda^{(2)}, \mathbf{w}, \gamma) = \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}(\|\mathbf{g}\|_2 + \gamma) + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i - \nu\|\tilde{\mathbf{x}}\|_1 + \sum_{i=n-k+1}^{n} (\lambda_i^{(1)} - \lambda_i^{(2)})\tilde{\mathbf{x}}_i + \sum_{i=1}^{n} (\lambda_i^{(1)} - \lambda_i^{(2)})\mathbf{w}_i - \gamma\sqrt{C_{\mathbf{w}}^2 + \sigma^2}.$$

(30)

The inner minimization over $\mathbf{w}$ is now doable. Setting the derivatives with respect to $\mathbf{w}_i$ to zero one obtains

$$\frac{\mathbf{w}(\|\mathbf{g}\|_2 + \gamma)}{\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}} + (\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}) = 0,$$

(31)

where $\lambda^{(1)} = [\lambda_1^{(1)}, \lambda_2^{(1)}, \ldots, \lambda_n^{(1)}]^T$, $\lambda^{(2)} = [\lambda_1^{(2)}, \lambda_2^{(2)}, \ldots, \lambda_n^{(2)}]^T$. From (31) one then has

$$\mathbf{w}(\|\mathbf{g}\|_2 + \gamma) = -\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}(\mathbf{h} + \lambda^{(1)} - \lambda^{(2)})$$

(32)

or in a norm form

$$\|\mathbf{w}\|_2^2(\|\mathbf{g}\|_2 + \gamma)^2 = (\|\mathbf{w}\|_2^2 + \sigma^2)\|\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}\|_2^2.$$

(33)

From (33) we then find

$$\|\mathbf{w}_{sol}\|_2 = \frac{\sigma\|\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}\|_2}{\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}\|_2^2}},$$

(34)

and from (32)

$$\mathbf{w}_{sol} = \frac{\sigma(\mathbf{h} + \lambda^{(1)} - \lambda^{(2)})}{\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}\|_2^2}}$$

(35)

where $\mathbf{w}_{sol}$ is of course the solution of the inner minimization over $\mathbf{w}$. Now, one should note that (34) and (35) are of course possible only if $\|\mathbf{g}\|_2 + \gamma - \|\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}\|_2 \geq 0$. Later in the paper we will recognize, that for $\lambda^{(1)}$ and $\lambda^{(2)}$ that are optimal in (29), validity of this condition essentially implies the regime (in (α, β) plane) where the worst-case $\|\mathbf{w}\|_2$ is finite with overwhelming probability (or equivalently, if for such $\lambda^{(1)}$ and $\lambda^{(2)}$ the condition is not valid then for the corresponding (α, β) the worst-case $\|\mathbf{w}\|_2$ is infinite with overwhelming probability). Plugging the value of $\mathbf{w}_{sol}$ from (35) back in (29) gives

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(1)}, \lambda^{(2)}, \gamma} \quad \sigma\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}\|_2^2} - \nu\|\tilde{\mathbf{x}}\|_1 + \sum_{i=n-k+1}^{n} (\lambda_i^{(1)} - \lambda_i^{(2)})\tilde{\mathbf{x}}_i - \gamma\sqrt{C_{\mathbf{w}}^2 + \sigma^2}$$

$$\text{subject to} \quad \lambda^{(i)} \geq 0, 1 \leq j \leq n, 1 \leq i \leq 2$$
$$\nu \geq 0$$
$$\nu - \lambda_i^{(1)} - \lambda_i^{(2)} = 0, 1 \leq i \leq n$$
$$\|\mathbf{g}\|_2 + \gamma - \|\mathbf{h} + \lambda^{(1)} - \lambda^{(2)}\|_2 \geq 0$$
$$\gamma \geq 0.$$

(36)

Let $\mathbf{z}^{(1)} = [1, 1, \ldots, 1]^T$. By plugging the constraint $\lambda^{(1)} = \nu\mathbf{z}^{(1)} - \lambda^{(2)}$ back into the objective function and making sure that $\nu - \lambda_i^{(2)} \geq 0, 1 \geq i \geq n$, one can remove $\lambda^{(1)}$ from the above optimization and get the

following

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}, \gamma} \quad \sigma\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - 2\lambda^{(2)}\|_2^2} - \nu\|\tilde{\mathbf{x}}\|_1 + \sum_{i=n-k+1}^{n} (\nu - 2\lambda_i^{(2)})\tilde{\mathbf{x}}_i - \gamma\sqrt{C_{\mathbf{w}}^2 + \sigma^2}$$

$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq \nu, 1 \leq i \leq n$$
$$\|\mathbf{g}\|_2 + \gamma - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - 2\lambda^{(2)}\|_2 \geq 0$$
$$\gamma \geq 0. \tag{37}$$

Since we assumed that $\tilde{\mathbf{x}}_i \geq 0, n - k + 1 \leq i \leq n$, and $\tilde{\mathbf{x}}_i = 0, 1 \leq i \leq n - k$ one then from (37) has

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}, \gamma} \quad \sigma\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - 2\lambda^{(2)}\|_2^2} - 2\sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i - \gamma\sqrt{C_{\mathbf{w}}^2 + \sigma^2}$$

$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq \nu, 1 \leq i \leq n$$
$$\|\mathbf{g}\|_2 + \gamma - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - 2\lambda^{(2)}\|_2 \geq 0$$
$$\gamma \geq 0. \tag{38}$$

After a simple scaling of $\lambda^{(2)}$ one finds that the following is an equivalent to (38)

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}, \gamma} \quad \sigma\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i - \gamma\sqrt{C_{\mathbf{w}}^2 + \sigma^2}$$

$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n$$
$$\|\mathbf{g}\|_2 + \gamma - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 \geq 0$$
$$\gamma \geq 0. \tag{39}$$

Now, the maximization over γ can be done. After setting the derivative to zero one finds

$$\frac{\|\mathbf{g}\|_2 + \gamma}{\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2}} - \sqrt{C_{\mathbf{w}}^2 + \sigma^2} = 0 \tag{40}$$

and after some algebra

$$\gamma_{opt} = \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}}\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 - \|\mathbf{g}\|_2, \tag{41}$$

where of course γ_{opt} would be the solution of (39) only if larger than or equal to zero. Alternatively of course $\gamma_{opt} = 0$. Now, based on these two scenarios we distinguish two different optimization problems:

11

1. *The "overwhelming" optimization*

$$\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}} \quad \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{42}$$

2. *The "non-overwhelming" optimization*

$$\xi_{nov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}} \quad \sqrt{C_{\mathbf{w}}^2 + \sigma^2}\|\mathbf{g}\|_2 - C_{\mathbf{w}}\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{43}$$

The "overwhelming" optimization is the equivalent to (39) if for its optimal values $\hat{\nu}$ and $\widehat{\lambda^{(2)}}$ holds

$$\sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}}\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 \leq \|\mathbf{g}\|_2, \tag{44}$$

We now summarize in the following lemma the results of this subsection.

Lemma 2. *Let $\hat{\nu}$ and $\widehat{\lambda^{(2)}}$ be the solutions of (42) and analogously let $\tilde{\nu}$ and $\widetilde{\lambda^{(2)}}$ be the solutions of (43). Let $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ be, as defined in (22), the optimal value of the objective function in (22). Then*

$$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \begin{cases} \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2} - \sum_{i=n-k+1}^{n} \widehat{\lambda_i^{(2)}} \tilde{\mathbf{x}}_i, & \text{if} \quad \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}}\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 \leq \|\mathbf{g}\|_2 \\ \sqrt{C_{\mathbf{w}}^2 + \sigma^2}\|\mathbf{g}\|_2 - C_{\mathbf{w}}\|\mathbf{h} + \tilde{\nu}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2)}}\|_2 - \sum_{i=n-k+1}^{n} \widetilde{\lambda_i^{(2)}} \tilde{\mathbf{x}}_i, & \text{otherwise} \end{cases}. \tag{45}$$

Moreover, let $\hat{\mathbf{w}}$ be the solution of (22). Then

$$\hat{\mathbf{w}}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \begin{cases} \frac{\sigma(\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}})}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}}, & \text{if} \quad \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}}\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 \leq \|\mathbf{g}\|_2 \\ \frac{C_{\mathbf{w}}(\mathbf{h} + \tilde{\nu}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2)}})}{\|\mathbf{h} + \tilde{\nu}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2)}}\|_2}, & \text{otherwise} \end{cases}, \tag{46}$$

and

$$\|\hat{\mathbf{w}}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})\|_2 = \begin{cases} \frac{\sigma\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}})\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}}, & \text{if} \quad \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}}\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 \leq \|\mathbf{g}\|_2 \\ C_{\mathbf{w}}, & \text{otherwise} \end{cases}. \tag{47}$$

Proof. The first part follows trivially. The second one follows from (35) by choosing the optimal $\hat{\nu}$ and $\widehat{\lambda^{(2)}}$ or alternatively $\tilde{\nu}$ and $\widetilde{\lambda^{(2)}}$. $\square$

2.1.2 Concentration of $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$

In this section we will show that $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ concentrates with high probability around its mean. To do so we will instead of looking at (45) look back at (22) which is the original definition of $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$. Now, before proceeding further we first recall on the following incredible result from [20] related to the concentrations of Lipschitz functions of Gaussian random variables.

Lemma 3 ([20, 54]). *Let $f_{lip}(\cdot) : R^n \longrightarrow R$ be a Lipschitz function such that $|f_{lip}(\mathbf{a}) - f_{lip}(\mathbf{b})| \leq c_{lip}\|\mathbf{a} - \mathbf{b}\|_2$. Let $\mathbf{a}$ be a vector comprised of i.i.d. zero-mean, unit variance Gaussian random variables and let $\epsilon_{lip} > 0$. Then*

$$P(|f_{lip}(\mathbf{a}) - Ef_{lip}(\mathbf{a})| \geq \epsilon_{lip}Ef_{lip}(\mathbf{a})) \leq \exp\left\{-\frac{(\epsilon_{lip}Ef_{lip}(\mathbf{a}))^2}{2c_{lip}^2}\right\}. \tag{48}$$

In the following lemma we will show that $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ is a Lipschitz function. To do so, we will, roughly speaking, assume that $\|\mathbf{w}\|_2$ in the definition of $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ is bounded by a large constants say $C_{\mathbf{w}}$. We recall here that our goal in this paper, though, is much bigger than creating a "constant type" bound on $\|\mathbf{w}\|_2$. Namely, we will actually establish the precise value that $\|\mathbf{w}\|_2$ takes in the worst case with overwhelming probability. Clearly, knowing that one could then use much better value than $C_{\mathbf{w}}$ to upper bound $\|\mathbf{w}\|_2$ in the definition of $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$. However, for the purposes of the concentration inequalities any constant (of course independent of n) is fine. In fact, any sub-root dependence on n would be fine too, it is just that in that case "overwhelming" wouldn't be negative exponential any more.

Lemma 4. *Let $\mathbf{g}$ and $\mathbf{h}$ be m and n dimensional vectors, respectively, with i.i.d. standard normal variables as their components. Let $\sigma > 0$ be an arbitrary scalar. Let $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ be as in (22). Further let $\epsilon_{lip} > 0$ be any constant. Then*

$$P(|\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) - E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})| \geq \epsilon_{lip}E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})) \leq \exp\left\{-\frac{(\epsilon_{lip}E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}))^2}{2(2C_{\mathbf{w}}^2 + \sigma^2)}\right\}. \tag{49}$$

Proof. We start by setting

$$f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) = \min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}^{(1)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(1)}\mathbf{w}_i\right). \tag{50}$$

Further, let $\mathbf{w}_{lip}^{(1)}$ be the solution of the minimization in (50). Then, clearly

$$f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) = \left(\sqrt{\|\mathbf{w}_{lip}^{(1)}\|_2^2 + \sigma^2}\|\mathbf{g}^{(1)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(1)}(\mathbf{w}_{lip}^{(1)})_i\right), \tag{51}$$

where $(\mathbf{w}_{lip}^{(1)})_i$ is the i-th index of $\mathbf{w}_{lip}^{(1)}$. In an analogous fashion set

$$f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) = \min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}^{(2)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(2)}\mathbf{w}_i\right), \tag{52}$$

13

and let $\mathbf{w}_{lip}^{(2)}$ be the solution of the minimization in (51). Then again clearly

$$f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) = \left(\sqrt{\|\mathbf{w}_{lip}^{(2)}\|_2^2 + \sigma^2} \|\mathbf{g}^{(2)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(2)} (\mathbf{w}_{lip}^{(2)})_i \right), \tag{53}$$

where of course $(\mathbf{w}_{lip}^{(2)})_i$ is the i-th index of $\mathbf{w}_{lip}^{(2)}$. Now assume that $f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) \neq f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)})$ (if they are equal we are trivially done). Further let $f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) < f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)})$ (the rest of the argument of course can trivially be flipped if $f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) > f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)})$). We then have

$$|f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) - f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)})| = f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) - f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)})$$

$$= \left(\sqrt{\|\mathbf{w}_{lip}^{(2)}\|_2^2 + \sigma^2} \|\mathbf{g}^{(2)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(2)} (\mathbf{w}_{lip}^{(2)})_i \right) - \left(\sqrt{\|\mathbf{w}_{lip}^{(1)}\|_2^2 + \sigma^2} \|\mathbf{g}^{(1)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(1)} (\mathbf{w}_{lip}^{(1)})_i \right)$$

$$\leq \left(\sqrt{\|\mathbf{w}_{lip}^{(1)}\|_2^2 + \sigma^2} \|\mathbf{g}^{(2)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(2)} (\mathbf{w}_{lip}^{(1)})_i \right) - \left(\sqrt{\|\mathbf{w}_{lip}^{(1)}\|_2^2 + \sigma^2} \|\mathbf{g}^{(1)}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i^{(1)} (\mathbf{w}_{lip}^{(1)})_i \right)$$

$$= \sqrt{\|\mathbf{w}_{lip}^{(1)}\|_2^2 + \sigma^2} (\|\mathbf{g}^{(2)}\|_2 - \|\mathbf{g}^{(1)}\|_2) + \sum_{i=1}^{n} (\mathbf{h}_i^{(2)} - \mathbf{h}_i^{(1)})(\mathbf{w}_{lip}^{(1)})_i$$

$$\leq \sqrt{\|\mathbf{w}_{lip}^{(1)}\|_2^2 + \sigma^2} (\|\mathbf{g}^{(2)} - \mathbf{g}^{(1)}\|_2) + \|(\mathbf{h}^{(2)} - \mathbf{h}^{(1)}\|_2 \|\mathbf{w}_{lip}^{(1)}\|_2$$

$$\leq \sqrt{2\|\mathbf{w}_{lip}^{(1)}\|_2^2 + \sigma^2} \sqrt{\|\mathbf{g}^{(2)} - \mathbf{g}^{(1)}\|_2^2 + \|\mathbf{h}^{(2)} - \mathbf{h}^{(1)}\|_2^2}$$

$$\leq \sqrt{2C_{\mathbf{w}}^2 + \sigma^2} \sqrt{\|\mathbf{g}^{(2)} - \mathbf{g}^{(1)}\|_2^2 + \|\mathbf{h}^{(2)} - \mathbf{h}^{(1)}\|_2^2}, \tag{54}$$

where the first inequality follows by sub-optimality of $\mathbf{w}_{lip}^{(1)}$ in (52). Connecting beginning and end in (54) and combining it with (50) one then has that $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ is Lipschitz with $c_{lip} = \sqrt{2C_{\mathbf{w}}^2 + \sigma^2}$. (49) then easily follows by Lemma 3. $\square$

One then has that $\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2$ and $\|\mathbf{h} + \tilde{\nu}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2)}}\|_2$ concentrate as well which automatically implies that $\hat{\mathbf{w}}$ also concentrates. More formally, one then has analogues to (49)

$$P(|\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 - E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2| \geq \epsilon_1^{(norm)} E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2) \leq e^{-\epsilon_2^{(norm)} n}$$

$$P(|\|\mathbf{h} + \tilde{\nu}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2)}}\|_2 - E\|\mathbf{h} + \tilde{\nu}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2)}}\|_2| \geq \epsilon_3^{(norm)} E\|\mathbf{h} + \tilde{\nu}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2)}}\|_2) \leq e^{-\epsilon_4^{(norm)} n}$$

$$P(|\|\hat{\mathbf{w}}\|_2 - E\|\hat{\mathbf{w}}\|_2| \geq \epsilon_1^{(\mathbf{w})} E\|\hat{\mathbf{w}}\|_2) \leq e^{-\epsilon_2^{(\mathbf{w})} n}, \tag{55}$$

where as usual $\epsilon_1^{(norm)} > 0$, $\epsilon_2^{(norm)} > 0$, and $\epsilon_1^{(\mathbf{w})} > 0$ are arbitrarily small constants and $\epsilon_3^{(norm)}$, $\epsilon_4^{(norm)}$, and $\epsilon_2^{(\mathbf{w})}$ are constant dependent on $\epsilon_1^{(norm)} > 0$, $\epsilon_2^{(norm)} > 0$, and $\epsilon_1^{(\mathbf{w})} > 0$, respectively, but independent of n.

Now, we return to the probabilistic analysis of (21). Combining (21), (22), and (49) we have

$$
\begin{aligned}
p_l &= P\left(\min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma,\tilde{\mathbf{x}},C_{\mathbf{w}})}\left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^n \mathbf{h}_i \mathbf{w}_i\right) + \mathbf{h}_{n+1}\sigma \geq \zeta_{obj}^{(l)}\right) \\
&= P\left(\xi(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) + \mathbf{h}_{n+1}\sigma \geq \zeta_{obj}^{(l)}\right) \\
&\geq \left(1 - \exp\left\{-\frac{(\epsilon_{lip} E\xi(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}))^2}{2(2C_{\mathbf{w}}^2 + \sigma^2)}\right\}\right) P\left((1-\epsilon_{lip})E\xi(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) + \mathbf{h}_{n+1}\sigma \geq \zeta_{obj}^{(l)}\right).
\end{aligned}
$$
(56)

Since $\mathbf{h}_{n+1}$ is a standard normal one easily has $P(\mathbf{h}_{n+1}\sigma \geq -\epsilon_1^{(\mathbf{h})}\sqrt{n}) \geq 1 - e^{-\epsilon_2^{(\mathbf{h})}n}$ where $\epsilon_1^{(\mathbf{h})} > 0$ is an arbitrarily small constant and $\epsilon_2^{(\mathbf{h})}$ is a constant dependent on $\epsilon_1^{(\mathbf{h})}$ and σ but independent on n. By choosing

$$
\zeta_{obj}^{(l)} = (1 - \epsilon_{lip})E\xi(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) - \epsilon_1^{(\mathbf{h})}\sqrt{n},
$$
(57)

one then from (56) has

$$
\begin{aligned}
p_l &= P\left(\min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma,\tilde{\mathbf{x}},C_{\mathbf{w}})}\left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^n \mathbf{h}_i \mathbf{w}_i\right) + \mathbf{h}_{n+1}\sigma \geq (1-\epsilon_{lip})E\xi(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) - \epsilon_1^{(\zeta)}\sqrt{n}\right) \\
&\geq \left(1 - \exp\left\{-\frac{(\epsilon_{lip} E\xi(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}))^2}{2(2C_{\mathbf{w}}^2 + \sigma^2)}\right\}\right)(1 - e^{-\epsilon_2^{(\mathbf{h})}n}).
\end{aligned}
$$
(58)

As stated after (20), (58) is conceptually enough to establish a "high probability" lower bound on ζ_{obj}. The next few steps that formally do so are rather obvious but we include them for the completeness. Combining (19) and (58) we obtain

$$
\begin{aligned}
P(\min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma,\tilde{\mathbf{x}},C_{\mathbf{w}})}(f_{obj}(\sigma,\mathbf{w}) + \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}g) \geq \zeta_{obj}^{(l)}) \\
\geq \left(1 - \exp\left\{-\frac{(\epsilon_{lip} E\xi(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}))^2}{2(2C_{\mathbf{w}}^2 + \sigma^2)}\right\}\right)(1 - e^{-\epsilon_2^{(\mathbf{h})}n}),,
\end{aligned}
$$
(59)

where $\zeta_{obj}^{(l)}$ is as in (57). Now, one further has

$$
\begin{aligned}
P(\min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma,\tilde{\mathbf{x}},C_{\mathbf{w}})}(f_{obj}(\sigma,\mathbf{w}) + \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}g) \geq \zeta_{obj}^{(l)}) \\
\leq P(\min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma,\tilde{\mathbf{x}},C_{\mathbf{w}})}(f_{obj}(\sigma,\mathbf{w})) + \sqrt{C_{\mathbf{w}}^2 + \sigma^2}g \geq \zeta_{obj}^{(l)}).
\end{aligned}
$$
(60)

Since g is a standard normal one easily again has $P(g\sqrt{C_{\mathbf{w}}^2 + \sigma^2} \leq \epsilon_1^{(g)}\sqrt{n}) \geq 1 - e^{-\epsilon_1^{(g)}n}$ where $\epsilon_1^{(g)} > 0$ is an arbitrarily small constant and $\epsilon_2^{(g)}$ is a constant dependent on $\epsilon_1^{(g)}$, σ, and $C_{\mathbf{w}}$ but independent on n. Applying this to the first term on the right hand side of the above inequality one obtains

$$
\begin{aligned}
P(\min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma,\tilde{\mathbf{x}},C_{\mathbf{w}})}(f_{obj}(\sigma,\mathbf{w})) + \sqrt{C_{\mathbf{w}}^2 + \sigma^2}g \geq \zeta_{obj}^{(l)}) \\
\leq P(\min_{[\mathbf{w}^T\sigma]^T \in S_{\mathbf{w}}(\sigma,\tilde{\mathbf{x}},C_{\mathbf{w}})}(f_{obj}(\sigma,\mathbf{w})) \geq \zeta_{obj}^{(l)} - \epsilon_1^{(g)}\sqrt{n}) + e^{-\epsilon_1^{(g)}n}.
\end{aligned}
$$
(61)

Now let $\zeta_{obj}^{lower} = \zeta_{obj}^l - \epsilon_1^{(g)}\sqrt{n}$. From (57) then obviously

$$\zeta_{obj}^{(lower)} = (1 - \epsilon_{lip})E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) - \epsilon_1^{(\mathbf{h})}\sqrt{n} - \epsilon_1^{(g)}\sqrt{n}. \tag{62}$$

Also let ϵ_{lower} be a constant such that

$$1 - e^{-\epsilon_{lower}n} \leq \left(1 - \exp\left\{-\frac{(\epsilon_{lip}E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}))^2}{2(2C_{\mathbf{w}}^2 + \sigma^2)}\right\}\right)(1 - e^{-\epsilon_2^{(\mathbf{h})}n}) - e^{-\epsilon_1^{(g)}n}. \tag{63}$$

Then a combination of (11), (59), (60), (61), (62), and (63) gives

$$P(\zeta_{obj} \geq \zeta_{obj}^{(lower)}) \geq P(\zeta_{obj}^{(help)} \geq \zeta_{obj}^{(lower)})(1 - e^{-\epsilon C_{\mathbf{w}}n})$$
$$= P(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})}(f_{obj}(\sigma, \mathbf{w})) \geq \zeta_{obj}^{(lower)}) \geq (1 - e^{-\epsilon_{lower}n})(1 - e^{-\epsilon C_{\mathbf{w}}n}). \tag{64}$$

We summarize the results from this subsection in the following lemma.

Lemma 5. *Let $\mathbf{v}$ be an $n \times 1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m \times n$ matrix of i.i.d. standard normal random variables. Consider an $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x} = \tilde{\mathbf{x}}$. Let then ζ_{obj} be as defined in (10) and let $\mathbf{w}$ be the solution of (14). Assume $P(\|\mathbf{w}\|_2 \leq C_{\mathbf{w}}) \geq 1 - e^{-\epsilon C_{\mathbf{w}}n}$ for an arbitrarily large constant $C_{\mathbf{w}}$ and a constant $\epsilon_{C_{\mathbf{w}}} > 0$ dependent on $C_{\mathbf{w}}$ but independent of n. Then there is a constant $\epsilon_{lower} > 0$*

$$P(\zeta_{obj} \geq \zeta_{obj}^{(lower)}) \geq (1 - e^{-\epsilon_{lower}n})(1 - e^{-\epsilon C_{\mathbf{w}}n}), \tag{65}$$

where

$$\zeta_{obj}^{(lower)} = (1 - \epsilon_{lip})E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) - \epsilon_1^{(\mathbf{h})}\sqrt{n} - \epsilon_1^{(g)}\sqrt{n}, \tag{66}$$

$\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ is as defined in (22), and $\epsilon_{lip}, \epsilon_1^{(\mathbf{h})}, \epsilon_1^{(g)}$ are all positive arbitrarily small constants.

Proof. Follows from the previous discussion. □

2.2 Upper-bounding ζ_{obj}

In this section we present a general framework for finding a "high-probability" upper bound on ζ_{obj}. To that end, let r and $C_{\mathbf{w}_{up}}$ be positive scalars (in this subsection we present a general framework and take these scalars to be arbitrary; however to make the bound as tight sa possible in the following subsection we will make them take particular values). Now, if we can show that there is a $\mathbf{w} \in R^n$ such that $\|\tilde{\mathbf{x}} + \mathbf{w}\|_1 \leq \|\tilde{\mathbf{x}}\|_1$ and $\|\mathbf{v} - A\mathbf{w}\|_2 \leq r$ with overwhelming probability then r can act as an upper bound on ζ_{obj}. We then start by looking at the following optimization problem

$$\begin{aligned}
\min_{\mathbf{w}} \quad & \|\tilde{\mathbf{x}} + \mathbf{w}\|_1 - \|\tilde{\mathbf{x}}\|_1 \\
& \left\|A_{\mathbf{v}}\begin{bmatrix}\mathbf{w}\\\sigma\end{bmatrix}\right\|_2 \leq r \\
& \|\mathbf{w}\|_2^2 \leq C_{\mathbf{w}_{up}}^2,
\end{aligned} \tag{67}$$

where $A_{\mathbf{v}}$ is as defined right after (14). If we can show that with overwhelming probability the objective value of the above optimization problem is negative then r will be a valid "high probability" upper-bound on ζ_{obj}. Moreover, it will be achieved by a $\mathbf{w}$ for which it will hold that $\|\mathbf{w}\|_2 \leq C_{\mathbf{w}_{up}}$.

We now proceed in a fashion similar to the one from Subsection 2.1.1. To remove the absolute values we introduce auxiliary variables $\mathbf{t}_i, 1 \leq i \leq n$, and transform the above problem to

$$
\begin{aligned}
\min_{\mathbf{w}, \mathbf{t}} \quad & \sum_{i=1}^{n} \mathbf{t}_i - \|\tilde{\mathbf{x}}\|_1 \\
\text{subject to} \quad & \tilde{\mathbf{x}}_i + \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n \\
& -\tilde{\mathbf{x}}_i - \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n \\
& \mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k \\
& -\mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k \\
& \left\| A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \right\|_2 \leq r \\
& \|\mathbf{w}\|_2^2 \leq C_{\mathbf{w}_{up}}^2.
\end{aligned}
\tag{68}
$$

We also slightly modify the first of the constraints from (67) in the following way

$$
\begin{aligned}
\min_{\mathbf{w}, \mathbf{t}, \mathbf{b}} \quad & \sum_{i=1}^{n} \mathbf{t}_i - \|\tilde{\mathbf{x}}\|_1 \\
\text{subject to} \quad & \tilde{\mathbf{x}}_i + \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n \\
& -\tilde{\mathbf{x}}_i - \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n \\
& \mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k \\
& -\mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k \\
& \|\mathbf{b}\|_2^2 \leq r^2 \\
& \begin{bmatrix} -A\mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} = b \\
& \|\mathbf{w}\|_2^2 \leq C_{\mathbf{w}_{up}}^2.
\end{aligned}
\tag{69}
$$

The Lagrange dual of the above problem then becomes

$$
\mathcal{L}(\lambda^{(1)}, \lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{t}, \mathbf{b}) = \sum_{i=1}^{n} \mathbf{t}_i - \|\tilde{\mathbf{x}}\|_1 + \sum_{i=n-k+1}^{n} \lambda_i^{(1)}(\tilde{\mathbf{x}}_i + \mathbf{w}_i - \mathbf{t}_i) + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}(-\tilde{\mathbf{x}}_i - \mathbf{w}_i - \mathbf{t}_i)
$$

$$
+ \sum_{i=1}^{n-k} \lambda_i^{(1)}(\mathbf{w}_i - \mathbf{t}_i) + \sum_{i=1}^{n-k} \lambda_i^{(2)}(-\mathbf{w}_i - \mathbf{t}_i) - \nu^{(1)} A\mathbf{w} + \nu^{(1)} \mathbf{v}\sigma - \nu^{(1)}\mathbf{b} + \gamma_1(\sum_{i=1}^{n} \mathbf{b}_1^2 - r^2) + \gamma_2(\|\mathbf{w}\|_2^2 - C_{\mathbf{w}_{up}}^2),
\tag{70}
$$

where $\nu^{(1)}$ is $1 \times m$ row vector of Lagrange variables and $\lambda^{(1)}, \lambda^{(2)}$ are as in previous sections. After rearranging terms we further have

$$
\mathcal{L}(\lambda^{(1)}, \lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{t}, \mathbf{b}) = -\|\tilde{\mathbf{x}}\|_1 + \sum_{i=1}^{n} \mathbf{t}_i(1 - \lambda_i^{(1)} - \lambda_i^{(2)}) + \sum_{i=n-k+1}^{n} \lambda_i^{(1)}(\tilde{\mathbf{x}}_i + \mathbf{w}_i)
$$

$$
+ \sum_{i=n-k+1}^{n} \lambda_i^{(2)}(-\tilde{\mathbf{x}}_i - \mathbf{w}_i) + \sum_{i=1}^{n-k} \lambda_i^{(1)}\mathbf{w}_i - \sum_{i=1}^{n-k} \lambda_i^{(2)}\mathbf{w}_i - \nu^{(1)} A\mathbf{w} + \nu^{(1)}\mathbf{v}\sigma - \nu^{(1)}\mathbf{b} + \gamma_1(\sum_{i=1}^{n} \mathbf{b}_1^2 - r^2) + \gamma_2(\|\mathbf{w}\|_2^2 - C_{\mathbf{w}_{up}}^2).
\tag{71}
$$

17

After a few further arrangements we finally have

$$\mathcal{L}(\lambda^{(1)}, \lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{t}, \mathbf{b}) = \sum_{i=1}^{n} \mathbf{t}_i (1 - \lambda_i^{(1)} - \lambda_i^{(2)}) + \sum_{i=n-k+1}^{n} (\lambda_i^{(1)} - \lambda_i^{(2)} - 1) \tilde{\mathbf{x}}_i$$

$$+ ((\lambda^{(1)} - \lambda^{(2)})^T - \nu^{(1)} A) \mathbf{w} + \nu^{(1)} \mathbf{v} \sigma - \nu^{(1)} \mathbf{b} + \gamma_1 (\sum_{i=1}^{n} \mathbf{b}_1^2 - r^2) + \gamma_2 (\|\mathbf{w}\|_2^2 - C_{\mathbf{w}_{up}}^2). \quad (72)$$

Setting $(\nu - \lambda_i^{(1)} - \lambda_i^{(2)}) = 0, 1 \leq i \leq n$, (to insure that the dual is bounded) we have

$$\mathcal{L}(\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{b}) = -2 \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$+ ((\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A) \mathbf{w} + \nu^{(1)} \mathbf{v} \sigma - \nu^{(1)} \mathbf{b} + \gamma_1 (\sum_{i=1}^{n} \mathbf{b}_1^2 - r^2) + \gamma_2 (\|\mathbf{w}\|_2^2 - C_{\mathbf{w}_{up}}^2). \quad (73)$$

Finally we can write a dual problem to (69)

$$\max_{\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2} \min_{\mathbf{w}, \mathbf{b}} \quad \mathcal{L}(\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{b})$$

$$\text{subject to} \quad 0 \leq \lambda_i^{(2)} \leq 1, 1 \leq i \leq n$$

$$\gamma_1 \geq 0,$$

$$\gamma_2 \geq 0, \quad (74)$$

where we of course use the fact that the strict duality obviously holds. Now, we minimize over $\mathbf{w}$ by setting the derivatives to zero

$$\frac{d\mathcal{L}(\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{b})}{d\mathbf{w}} = ((\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A)^T + 2\gamma_2 \mathbf{w}. \quad (75)$$

From (75) we easily have

$$\mathbf{w} = \frac{((\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A)^T}{2\gamma_2}. \quad (76)$$

Plugging (75) back in (73) we further have

$$\mathcal{L}(\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{b}) = -2 \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$- \frac{\|(\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A\|_2}{4\gamma_2} + \nu^{(1)} \mathbf{v} \sigma - \nu^{(1)} \mathbf{b} + \gamma_1 (\sum_{i=1}^{n} \mathbf{b}_1^2 - r^2) - \gamma_2 C_{\mathbf{w}_{up}}^2. \quad (77)$$

Now, we minimize $\mathcal{L}(\lambda^{(1)}, \lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{b})$ over $\mathbf{b}$ by setting the derivatives to zero

$$\frac{d\mathcal{L}(\lambda^{(1)}, \lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{b})}{d\mathbf{b}} = -\nu^{(1)} + 2\gamma_1 \mathbf{w}. \quad (78)$$

From (78) we easily have

$$\mathbf{b} = \frac{\nu^{(1)}}{2\gamma_1}. \quad (79)$$

18

Plugging (79) back in (77) we have

$$\mathcal{L}(\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2) = -2 \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i + -\frac{\|(\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A\|_2}{2\gamma_2} + \nu^{(1)} \mathbf{v}\sigma - \frac{\|\nu^{(1)}\|_2}{4\gamma_1} - \gamma_1 r^2 - \gamma_2 C_{\mathbf{w}_{up}}^2,$$

(80)

and finally an equivalent to (69)

$$\max_{\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2} \quad \mathcal{L}(\lambda^{(1)}, \lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2)$$

$$\text{subject to} \quad 0 \le \lambda_i^{(2)} \le 1, 1 \le i \le n$$

$$\gamma_1 \ge 0,$$

$$\gamma_2 \ge 0. \tag{81}$$

After doing the trivial maximization over γ_1 and γ_2 one obtains

$$\max_{\lambda^{(2)}, \nu^{(1)}} \quad -2 \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i - C_{\mathbf{w}_{up}} \|(\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A\|_2 + \nu^{(1)} \mathbf{v}\sigma - \|\nu^{(1)}\|_2 r$$

$$\text{subject to} \quad 0 \le \lambda_i^{(2)} \le 1, 1 \le i \le n. \tag{82}$$

We rewrite (82) in a slightly more convenient form

$$-\min_{\lambda^{(2)}, \nu^{(1)}} \max_{\|\mathbf{a}\|_2 = C_{\mathbf{w}_{up}}} \quad ((\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A)\mathbf{a} - \nu^{(1)} \mathbf{v}\sigma + \|\nu^{(1)}\|_2 r + 2 \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad 0 \le \lambda_i^{(2)} \le 1, 1 \le i \le n. \tag{83}$$

Now let us define $f_{obj}^{(up)}$ as

$$-f_{obj}^{(up)} = -\min_{\lambda^{(2)}, \nu^{(1)}} \max_{\|\mathbf{a}\|_2 = C_{\mathbf{w}_{up}}} \quad ((\mathbf{z}^{(1)} - 2\lambda^{(2)})^T - \nu^{(1)} A)\mathbf{a} - \nu^{(1)} \mathbf{v}\sigma + \|\nu^{(1)}\|_2 r + 2 \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad 0 \le \lambda_i^{(2)} \le 1, 1 \le i \le n. \tag{84}$$

Any r such that $\lim_{n\to} P(f_{obj}^{(up)} \ge 0) = 1$ is then a valid "high-probability" upper bound.

We now introduce a refinement of a lemma from [62] which itself is a slightly modified Lemma 1 (Lemma 1 is of course the backbone of the escape through a mesh theorem utilized in [63]).

Lemma 6. *Let A be an $m \times n$ matrix with i.i.d. standard normal components. Let $\mathbf{g}$ and $\mathbf{h}$ be $m \times 1$ and $(n+1) \times 1$ vectors, respectively, with i.i.d. standard normal components. Also, let g be a standard normal random variable and let Λ be a set such that $\Lambda = (\lambda^{(2)} | 0 \le \lambda_i^{(2)} \le 1, 1 \le i \le n)$. Then*

$$P\left(\min_{\lambda^{(2)} \in \Lambda, \nu^{(1)} \in R^n \backslash 0} \max_{\|\mathbf{a}\|_2 = C_{\mathbf{w}_{up}}} \left(-\nu^{(1)} \begin{bmatrix} A & \mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \sigma \end{bmatrix} + \|\nu^{(1)}\|_2 g - \psi_{\mathbf{a}, \lambda^{(2)}, \nu^{(1)}}\right) \ge 0\right)$$

$$\ge P\left(\min_{\lambda^{(2)} \in \Lambda, \nu^{(1)} \in R^n \backslash 0} \max_{\|\mathbf{a}\|_2 = 1} \left(\|\nu^{(1)}\|_2 \left(\sum_{i=1}^{n} \mathbf{h}_i \mathbf{a}_i + \mathbf{h}_{n+1}\sigma\right) + \sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2} \sum_{i=1}^{m} \mathbf{g}_i \nu_i^{(1)} - \psi_{\mathbf{a}, \lambda^{(2)}, \nu^{(1)}}\right) \ge 0\right).$$

(85)

19

Let

$$\psi_{\mathbf{a},\lambda^{(2)},\nu^{(1)}} = \epsilon_3^{(g)}\sqrt{n}\|\nu^{(1)}\|_2 - \mathbf{a}^T(\mathbf{z}^{(1)} - 2\lambda^{(2)}) - \|\nu^{(1)}\|_2 r - 2\sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i, \qquad (86)$$

with $\epsilon_3^{(g)} > 0$ being an arbitrarily small constant independent of n. The left-hand side of the inequality in (85) is then the following probability of interest

$$p_u = P\Big(\min_{\lambda^{(2)}\in\Lambda,\nu^{(1)}\in R^n\backslash 0}\max_{\|\mathbf{a}\|_2=C_{\mathbf{w}_{up}}}(\|\nu^{(1)}\|_2(\sum_{i=1}^{n}\mathbf{h}_i\mathbf{a}_i + \mathbf{h}_{n+1}\sigma) + \sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\sum_{i=1}^{m}\mathbf{g}_i\nu_i^{(1)}$$

$$- \epsilon_3^{(g)}\sqrt{n}\|\nu^{(1)}\|_2 + \mathbf{a}^T(\mathbf{z}^{(1)} - 2\lambda^{(2)}) + \|\nu^{(1)}\|_2 r + 2\sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i) \geq 0\Big).$$

After solving the inner maximization over $\mathbf{a}$ and pulling out $\|\nu\|_2$ one has

$$p_u = P\Big(\min_{\lambda^{(2)}\in\Lambda,\nu\in R^n\backslash 0}(C_{\mathbf{w}_{up}}\|\mathbf{h} + \frac{1}{\|\nu^{(1)}\|_2}(\mathbf{z}^{(1)} - 2\lambda^{(2)})\|_2 + \mathbf{h}_{n+1}\sigma$$

$$+ \sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\sum_{i=1}^{m}\mathbf{g}_i\frac{\nu_i^{(1)}}{\|\nu^{(1)}\|_2} - \epsilon_3^{(g)}\sqrt{n} + r + 2\sum_{i=n-k+1}^{n}\frac{\lambda_i^{(2)}}{\|\nu^{(1)}\|_2}\tilde{\mathbf{x}}_i) \geq 0\Big).$$

After minimization of the second term over a unit norm vector we further have

$$p_u = P\Big(\min_{\lambda^{(2)}\in\Lambda,\nu\in R^n\backslash 0}(C_{\mathbf{w}_{up}}\|\mathbf{h} + \frac{1}{\|\nu^{(1)}\|_2}(\mathbf{z}^{(1)} - 2\lambda^{(2)})\|_2 + \mathbf{h}_{n+1}\sigma$$

$$- \sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}\| - 2 - \epsilon_3^{(g)}\sqrt{n} + r + 2\sum_{i=n-k+1}^{n}\frac{\lambda_i^{(2)}}{\|\nu^{(1)}\|_2}\tilde{\mathbf{x}}_i) \geq 0\Big). \quad (87)$$

Now we change variables so that $\nu = \frac{1}{\|\nu^{(1)}\|_2}$ and $\lambda^{(2)} = \frac{2\lambda^{(2)}}{\|\nu^{(1)}\|_2}$ and redefine Λ by setting

$$\Lambda^{(2)} = \{\lambda^{(2)} \in R^n | 0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n\}. \qquad (88)$$

We also recall that $\mathbf{z}^{(1)}$ remains as defined right after (36). Plugging all of this back in (87) gives us

$$p_u = P\Big(r + \mathbf{h}_{n+1}\sigma - \epsilon_3^{(g)}\sqrt{n} - \max_{\lambda^{(2)}\in\Lambda^{(2)},\nu\geq 0}(\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i) \geq 0\Big).$$

$$(89)$$

Now, let

$$\xi_{up}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},C_{\mathbf{w}_{up}}) = \max_{\lambda^{(2)}\in\Lambda^{(2)},\nu\geq 0}(\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i).$$

$$(90)$$

In the following lemma we will show that $\xi_{up}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},C_{\mathbf{w}_{up}})$ is a Lipschitz function.

Lemma 7. *Let $\mathbf{g}$ and $\mathbf{h}$ be m and n dimensional vectors, respectively, with i.i.d. standard normal variables as their components. Let $\sigma > 0$ be an arbitrary scalar. Let $\xi_{up}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},C_{\mathbf{w}_{up}})$ be as in (90). Further let*

$\epsilon_{lip} > 0$ *be any constant. Then*

$$P(|\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}}) - E\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}})| \geq \epsilon_{lip} E\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}})) \leq \exp\left\{-\frac{(\epsilon_{lip} E\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}}))^2}{2(2C_{\mathbf{w}}^2 + \sigma^2)}\right\}.$$
$$(91)$$

Proof. The proof will be similar to the corresponding one from Subsection 2.1.2. We start by setting

$$f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) = \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0} (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}^{(1)}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h}^{(1)} + \nu\mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i).$$
$$(92)$$

Further, let $\nu^{(lip_1)}$ and $\lambda^{(lip_1)}$ be the solutions of the minimization in (92). Then, clearly

$$f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) = (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}^{(1)}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h}^{(1)} + \nu^{(lip_1)}\mathbf{z}^{(1)} - \lambda^{(lip_1)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(lip_1)}\tilde{\mathbf{x}}_i). \quad (93)$$

In an analogous fashion set

$$f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) = \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0} (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}^{(2)}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h}^{(2)} + \nu\mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i),$$
$$(94)$$

and let $\nu^{(lip_2)}$ and $\lambda^{(lip_2)}$ be the solutions of the minimization in (94). Then, clearly

$$f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) = (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}^{(2)}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h}^{(2)} + \nu^{(lip_2)}\mathbf{z}^{(1)} - \lambda^{(lip_2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(lip_2)}\tilde{\mathbf{x}}_i), \quad (95)$$

Now assume that $f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) \neq f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)})$ (if they are equal we are trivially done). Further let $f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) < f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)})$ (the rest of the argument of course can trivially be flipped if $f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)}) >$

$f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}))$. We then have

$$
|f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) - f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)})| = f_{lip}(\mathbf{g}^{(2)}, \mathbf{h}^{(2)}) - f_{lip}(\mathbf{g}^{(1)}, \mathbf{h}^{(1)})
$$

$$
= (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2} \|\mathbf{g}^{(2)}\|_2 - C_{\mathbf{w}_{up}} \|\mathbf{h}^{(2)} + \nu^{(lip_2)} \mathbf{z}^{(1)} - \lambda^{(lip_2)}\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(lip_2)} \tilde{\mathbf{x}}_i)
$$

$$
- (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2} \|\mathbf{g}^{(1)}\|_2 - C_{\mathbf{w}_{up}} \|\mathbf{h}^{(1)} + \nu^{(lip_1)} \mathbf{z}^{(1)} - \lambda^{(lip_1)}\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(lip_1)} \tilde{\mathbf{x}}_i)
$$

$$
\leq (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2} \|\mathbf{g}^{(2)}\|_2 - C_{\mathbf{w}_{up}} \|\mathbf{h}^{(2)} + \nu^{(lip_1)} \mathbf{z}^{(1)} - \lambda^{(lip_1)}\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(lip_1)} \tilde{\mathbf{x}}_i)
$$

$$
- (\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2} \|\mathbf{g}^{(1)}\|_2 - C_{\mathbf{w}_{up}} \|\mathbf{h}^{(1)} + \nu^{(lip_1)} \mathbf{z}^{(1)} - \lambda^{(lip_1)}\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(lip_1)} \tilde{\mathbf{x}}_i)
$$

$$
= \sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}(\|\mathbf{g}^{(2)}\|_2 - \|\mathbf{g}^{(1)}\|_2) - C_{\mathbf{w}_{up}}(\|\mathbf{h}^{(2)} + \nu^{(lip_1)} \mathbf{z}^{(1)} - \lambda^{(lip_1)}\|_2 - \|\mathbf{h}^{(2)} + \nu^{(lip_1)} \mathbf{z}^{(1)} - \lambda^{(lip_1)}\|_2)
$$

$$
\leq \sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2} \|\mathbf{g}^{(2)} - \mathbf{g}^{(1)}\|_2 + C_{\mathbf{w}_{up}}(\|\mathbf{h}^{(2)} - \mathbf{h}^{(1)}\|_2)
$$

$$
\leq \sqrt{2C_{\mathbf{w}_{up}}^2 + \sigma^2} \sqrt{\|\mathbf{g}^{(2)} - \mathbf{g}^{(1)}\|_2^2 + (\|\mathbf{h}^{(2)} - \mathbf{h}^{(1)}\|_2^2)}, \quad (96)
$$

where the first inequality follows by sub-optimality of ν^{lip_1} and $\lambda^{(lip_1)}$ in (94). Connecting beginning and end in (96) and combining it with (92) one then has that $\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}})$ is Lipschitz with $c_{lip} = \sqrt{2C_{\mathbf{w}}^2 + \sigma^2}$. (91) then easily follows by Lemma 3. $\qquad \square$

We continue by following the line of arguments right after (56). As stated there $P(\mathbf{h}_{n+1}\sigma \geq -\epsilon_1^{(\mathbf{h})}\sqrt{n}) \geq 1 - e^{-\epsilon_2^{(\mathbf{h})}n}$ where $\epsilon_1^{(\mathbf{h})} > 0$ is an arbitrarily small constant and $\epsilon_2^{(\mathbf{h})}$ is a constant dependent on $\epsilon_1^{(\mathbf{h})}$ and σ but independent on n. Set

$$
r = \zeta_{obj}^{(u)} = (1 + \epsilon_{lip})E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}}) + \epsilon_1^{(\mathbf{h})}\sqrt{n} + \epsilon_3^{(g)}\sqrt{n}. \quad (97)
$$

One then has after combing (89) and Lemma 7

$$
p_u = P(r + \mathbf{h}_{n+1}\sigma - \epsilon_3^{(g)}\sqrt{n} - \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0}(\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i) \geq 0)
$$

$$
\geq P((1 + \epsilon_{lip})E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}})
$$

$$
\geq \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0}(\sqrt{C_{\mathbf{w}_{up}}^2 + \sigma^2}\|\mathbf{g}\|_2 - C_{\mathbf{w}_{up}}\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i))(1 - e^{-\epsilon_2^{(\mathbf{h})}n})
$$

$$
\geq \left(1 - \exp\left\{-\frac{(\epsilon_{lip}E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}}))^2}{2(2C_{\mathbf{w}_{up}}^2 + \sigma^2)}\right\}\right)(1 - e^{-\epsilon_2^{(\mathbf{h})}n}). \quad (98)
$$

As stated after (20), (98) is conceptually enough to establish a "high probability" upper bound on ζ_{obj}. What is left is to connect it with (84). Combining (98), (85), and (84) we then obtain

$$
P(f_{obj}^{(up)} \geq 0) \geq \left(1 - \exp\left\{-\frac{(\epsilon_{lip}E\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}}))^2}{2(2C_{\mathbf{w}_{up}}^2 + \sigma^2)}\right\}\right)(1 - e^{-\epsilon_2^{(\mathbf{h})}n})(1 - e^{-\epsilon_4^{(g)}n}), \quad (99)
$$

where we used the fact that g is the standard normal and therefore $P(g - \epsilon_3^{(g)}\sqrt{n} \leq 0) \geq (1 - e^{-\epsilon_4^{(g)}n})$ for an arbitrarily small $\epsilon_3^{(g)} > 0$ and a constant $\epsilon_4^{(g)}$ dependent on $\epsilon_3^{(g)}$ but independent of n.

We are now in position to summarize results from this subsection in the following lemma which is essentially an "upper-bound" analogue to Lemma 5.

Lemma 8. *Let $\mathbf{v}$ be an $n \times 1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m \times n$ matrix of i.i.d. standard normal random variables. Consider an $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x} = \tilde{\mathbf{x}}$. Let then ζ_{obj} be as defined in (10) and let $\mathbf{w}$ be the solution of (14). There is a constant $\epsilon_{upper} > 0$*

$$P(\zeta_{obj} \leq \zeta_{obj}^{(upper)}) \geq 1 - e^{-\epsilon_{upper}n}, \tag{100}$$

where

$$\zeta_{obj}^{(upper)} = (1 + \epsilon_{lip})E\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}}) + \epsilon_1^{(\mathbf{h})}\sqrt{n} + \epsilon_3^{(g)}\sqrt{n}, \tag{101}$$

$\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}})$ is as defined in (90), $\epsilon_{lip}, \epsilon_1^{(\mathbf{h})}, \epsilon_3^{(g)}$ are all positive arbitrarily small constants, and $C_{\mathbf{w}_{up}}$ is a constant such that $\|\mathbf{w}\|_2 \leq C_{\mathbf{w}_{up}}$.

Proof. Follows from the previous discussion. $\square$

2.3 Matching upper and lower bounds

In this section we specialize the general bounds introduced above and show how they can match each other. We will divide presentation in three subsections. In the first of the subsections we will make a connection to the noiseless case and show how one can then remove the constraint from (45), (46), and (47). In the second subsection we will consider a $\mathbf{w}$ such that $|\|\mathbf{w}\|_2 - \|\hat{\mathbf{w}}\|_2| \geq \epsilon_{\mathbf{w}_{up}}\|\hat{\mathbf{w}}\|_2$. We will then quantify how much the lower bound that can be computed for such a $\mathbf{w}$ through the framework presented in Section 2.1 deviates from the optimal one obtained for $\hat{\mathbf{w}}$. In the last subsection we will then show that there will be a $\mathbf{w}$ such that the upper bound computed through the framework presented in Section 2.2 will deviate less. That will in essence establish that upper and lower bounds computed in the previous sections indeed match. We will then draw conclusions as for the consequences which such a matching of the bounds leaves on a couple of LASSO parameters.

2.3.1 Connection to the ℓ_1 optimization

In this subsection we establish a connection between the constraint in (45), (46), and (47) and the fundamental performance characterization of ℓ_1 optimization derived in [62] (and of course earlier in the context of neighborly polytopes in [25]). We first recall on the condition from Lemma 2. The condition states

$$\sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}}\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 \leq \|\mathbf{g}\|_2, \tag{102}$$

where $C_{\mathbf{w}}$ is an arbitrarily large constant and $\hat{\nu}$ and $\widehat{\lambda^{(2)}}$ are the solutions of

$$\max \quad \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad 0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n$$

$$\nu \geq 0. \tag{103}$$

Now we note the following equivalent to (103) for the case when nonzero components of $\tilde{\mathbf{x}}$ are infinite

$$
\begin{aligned}
\max \quad & \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} \\
\text{subject to} \quad & 0 \le \lambda_i^{(2)} \le 2\nu, 1 \le i \le n - k \\
& \lambda_i^{(2)} = 0, n - k + 1 \le i \le n \\
& \nu \ge 0.
\end{aligned} \tag{104}
$$

Now, to make the new observations easily comparable to the corresponding ones from [61, 63] we set

$$
\bar{\mathbf{h}} = [|\mathbf{h}|_{(1)}^{(1)}, |\mathbf{h}|_{(2)}^{(2)}, \dots, |\mathbf{h}|_{(n-k)}^{(n-k)}, \mathbf{h}_{n-k+1}, \mathbf{h}_{n-k+2}, \dots, \mathbf{h}_n]^T, \tag{105}
$$

where $[|\mathbf{h}|_{(1)}^{(1)}, |\mathbf{h}|_{(2)}^{(2)}, \dots, |\mathbf{h}|_{(n-k)}^{(n-k)}]$ are magnitudes of $[\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_{n-k}]$ sorted in increasing order (possible ties in the sorting process are of course broken arbitrarily). Also we let $\mathbf{z}^{(2)}$ be such that $\mathbf{z}_i^{(2)} = -\mathbf{z}_i^{(1)}, n - k + 1 \le i \le n$ and $\mathbf{z}_i^{(2)} = \mathbf{z}_i^{(1)}, 1 \le i \le n - k$. It is then relatively easy to see that the above optimization problem is equivalent to

$$
\begin{aligned}
\max \quad & \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\bar{\mathbf{h}} - \nu\mathbf{z}^{(2)} + \lambda^{(2)}\|_2^2} \\
\text{subject to} \quad & 0 \le \lambda_i^{(2)} \le \nu, 1 \le i \le n - k \\
& \lambda_i^{(2)} = 0, n - k + 1 \le i \le n \\
& \nu \ge 0.
\end{aligned} \tag{106}
$$

Let ν_{ℓ_1} and $\lambda^{(\ell_1)}$ be the solution of the above maximization. Then, as we showed in [63] and [62], the inequality

$$
E\|\mathbf{g}\|_2 > E\|\bar{\mathbf{h}} - \nu_{\ell_1}\mathbf{z}^{(2)} + \lambda^{(\ell_1)}\|_2 \tag{107}
$$

establishes the following fundamental performance characterization of the ℓ_1 optimization algorithm from (2) that could be used instead of LASSO to recover $\mathbf{x}$ in (1) (which is a noiseless version of (3))

$$
(1 - \beta_w)\frac{\sqrt{\frac{2}{\pi}}e^{-(\text{erfinv}(\frac{1-\alpha_w}{1-\beta_w}))^2}}{\alpha_w} - \sqrt{2}\text{erfinv}(\frac{1-\alpha_w}{1-\beta_w}) = 0, \tag{108}
$$

where of course $\alpha_w = \frac{m}{n}$ and $\beta_w = \frac{k}{n}$. As it is also shown in [63] and [62] both of the quantities under the expected values in (107) nicely concentrate. Then with overwhelming probability one has that for any pair (α, β) that satisfies (or lies below) the above fundamental performance characterization of ℓ_1 optimization

$$
\|\mathbf{g}\|_2 > \|\bar{\mathbf{h}} - \nu_{\ell_1}\mathbf{z}^{(2)} + \lambda^{(\ell_1)}\|_2. \tag{109}
$$

Moreover, since $\lambda_i^{(2)} \ge 0, n - k + 1 \le i \le n$, in (103) one actually has that (109) implies that with overwhelming probability

$$
\|\mathbf{g}\|_2 > \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2, \tag{110}
$$

which for sufficiently large $C_{\mathbf{w}}$ is the same as (102). We then in what follows assume that pair (α, β) is such that it satisfies the fundamental ℓ_1 optimization performance characterization (or is in the region below it) and therefore proceed by ignoring the condition (102). (Strictly speaking, all our overwhelming probabilities below should be multiplied with an overwhelming probability that (108) holds; to maintain writing easier we will skip this detail.)

2.3.2 Deviation from the lower-bound

In this subsection we show that $\|\mathbf{w}_{lasso}\|_2$ can not deviate substantially from $\|\hat{\mathbf{w}}\|_2$ without substantially affecting the value of the lower bound on the objective in (6) that is derived in Section 2.1. To that end let us assume that there is a $\mathbf{w}_{off}$ that is the solution of the LASSO from (6) (or to be slightly more precise that is such that $\hat{\mathbf{x}} = \tilde{\mathbf{x}} + \mathbf{w}_{off}$, where obviously $\hat{\mathbf{x}}$ is the solution of (6)). Further, let $\|\|\mathbf{w}_{off}\|_2 - \|\hat{\mathbf{w}}\|_2\| \geq \epsilon_{\mathbf{w}_{up}}\|\hat{\mathbf{w}}\|_2$, where $\epsilon_{\mathbf{w}_{up}}$ is an arbitrarily small constant.

One can then proceed by repeating the same line of thought as in Section 2.1. The only difference will be that now $C_{\mathbf{w}} = \|\mathbf{w}_{off}\|_2$ and consequently in the definition of $S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})$, $\|\mathbf{w}\|_2 \leq C_{\mathbf{w}}$ changes to $\|\mathbf{w}\|_2 = C_{\mathbf{w}} = \|\mathbf{w}_{off}\|_2$. This difference will of course not affect the concept presented in Section 2.1. The only real consequence will be the change of (22). Adapted to the new scenario (22) becomes

$$\xi_{off}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) = \min_{\mathbf{w}, \mathbf{t}} \quad \sqrt{\|\mathbf{w}_{off}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i$$

$$\text{subject to} \quad \sum_{i=1}^{n} \mathbf{t}_i \leq \|\tilde{\mathbf{x}}\|_1$$
$$\tilde{\mathbf{x}}_i + \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n$$
$$-\tilde{\mathbf{x}}_i - \mathbf{w}_i - \mathbf{t}_i \leq 0, n - k + 1 \leq i \leq n$$
$$\mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k$$
$$-\mathbf{w}_i - \mathbf{t}_i \leq 0, 1 \leq i \leq n - k$$
$$\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \leq \sqrt{\mathbf{w}_{off}^2 + \sigma^2}. \tag{111}$$

One can then proceed further with solving the Lagrangian to obtain

$$\xi_{off}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) = \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0} (\sqrt{\mathbf{w}_{off}^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{w}_{off}\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i). \tag{112}$$

Using the probabilistic arguments from Section 2.1 one then from Lemma 5 has that if $\mathbf{w}_{off}$ is the solution of (6) then its objective value with overwhelming probability is lower bounded by $(1 - \epsilon_{lip}) E\xi_{off}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off})$ ($\xi_{off}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off})$ is structurally the same as $\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, C_{\mathbf{w}_{up}})$ from (90) and therefore easily concentrates based on Lemma 7). We will now consider in parallel the following lower bound from (42) (clearly, choosing $\mathbf{w}_{off} = \hat{\mathbf{w}}$ would make (112) equivalent to (42)).

$$\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu \geq 0, \lambda^{(2)} \in \Lambda^{(2)}} \sigma \sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i. \tag{113}$$

Now, let as usual $\hat{\nu}$ and $\widehat{\lambda^{(2)}}$ be the solutions of (113). Let

$$\xi_{help}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) = \sqrt{\mathbf{w}_{off}^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{w}_{off}\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 - \sum_{i=n-k+1}^{n} \widehat{\lambda_i^{(2)}} \tilde{\mathbf{x}}_i. \tag{114}$$

Then

$$\xi_{off}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) - \xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \geq \xi_{help}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) - \xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$$
$$= \sqrt{\mathbf{w}_{off}^2 + \sigma^2}\|\mathbf{g}\|_2 - \mathbf{w}_{off}\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 - \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}. \quad (115)$$

For the simplicity let $\|\|\mathbf{w}_{off}\|_2 - \|\hat{\mathbf{w}}\|_2\| = \epsilon_{\mathbf{w}_{up}}\|\hat{\mathbf{w}}\|_2$ (this restriction is clearly more conservative than $\|\|\mathbf{w}_{off}\|_2 - \|\hat{\mathbf{w}}\|_2\| \geq \epsilon_{\mathbf{w}_{up}}\|\hat{\mathbf{w}}\|_2$). Now, we switch to expectations and ignore all ϵ except $\epsilon_{\mathbf{w}_{up}}$. Since every quantity that we will consider (see (55)) concentrates ϵ's in concentration inequalities can be made arbitrarily close to zero; moreover once $\epsilon_{\mathbf{w}_{up}}$ is fixed all other ϵ's can be made arbitrarily small compared to $\epsilon_{\mathbf{w}_{up}}$. Also, we will show derivation for $\mathbf{w}_{off} = (1 + \epsilon_{\mathbf{w}_{up}})\|\hat{\mathbf{w}}\|_2$ (the derivation for the case $\mathbf{w}_{off} = (1 - \epsilon_{\mathbf{w}_{up}})\|\hat{\mathbf{w}}\|_2$ is completely analogous).

Now, to facilitate writing we then set all ϵ's except $\epsilon_{\mathbf{w}_{up}}$ to zero. We then have

$$E\xi_{off}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) - E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \geq E\xi_{help}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) - E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$$
$$\doteq \sqrt{(1 + \epsilon_{\mathbf{w}_{up}})^2(E\|\hat{\mathbf{w}}\|_2)^2 + \sigma^2}E\|\mathbf{g}\|_2 - (1 + \epsilon_{\mathbf{w}_{up}})E\|\hat{\mathbf{w}}\|_2 E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2$$
$$- \sigma\sqrt{(E\|\mathbf{g}\|_2)^2 - (E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2)^2} \quad (116)$$

where $\doteq$ means that equality is not exact but for a fixed $\epsilon_{\mathbf{w}_{up}}$ can be made as close to it as needed. In a similar fashion we have

$$E\|\hat{\mathbf{w}}\|_2 \doteq \frac{\sigma E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2}{\sqrt{(E\|\mathbf{g}\|_2)^2 - (E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2)^2}}. \quad (117)$$

Before we proceed further we simplify the notation with the following change of variables.

$$\begin{aligned}
g_E &= E\|\mathbf{g}\|_2 \\
h_E &= E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 \\
\xi_E &= \sigma\sqrt{(E\|\mathbf{g}\|_2)^2 - (E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2)^2} = \sigma\sqrt{g_E^2 - h_E^2} \\
w_E &= E\|\hat{\mathbf{w}}\|_2 = \frac{\sigma h_E}{\sqrt{g_E^2 - h_E^2}}.
\end{aligned} \quad (118)$$

From (118) one easily has

$$\begin{aligned}
h_E^2 &= g_E^2 - \frac{\xi_E^2}{\sigma^2} \\
w_E &= \frac{h_E\sigma^2}{\xi_E}.
\end{aligned} \quad (119)$$

Then a combination of (116), (118), and (119) gives

$$E\xi_{help}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) - E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \doteq \sqrt{(1 + \epsilon_{\mathbf{w}_{up}})^2 (\frac{h_E^2 \sigma^4}{\xi_E^2}) + \sigma^2 g_E} - (1 + \epsilon_{\mathbf{w}_{up}}) w_E h_E - \xi_E$$

$$= (1 + \epsilon_{\mathbf{w}_{up}}) \frac{g_E^2 \sigma^2}{\xi_E} \sqrt{1 - \frac{\xi_E^2 (2\epsilon_{\mathbf{w}_{up}} + \epsilon_{\mathbf{w}_{up}}^2)}{(1 + \epsilon_{\mathbf{w}_{up}})^2 g_E^2 \sigma^2}} - (1 + \epsilon_{\mathbf{w}_{up}}) \frac{g_E^2 \sigma^2}{\xi_E} + \epsilon_{\mathbf{w}_{up}} \xi_E. \quad (120)$$

Now, assuming that $\epsilon_{\mathbf{w}_{up}}$ is small (and recognizing that $\xi_E \leq g_E \sigma$) from (120) we have

$$(1 + \epsilon_{\mathbf{w}_{up}}) \frac{g_E^2 \sigma^2}{\xi_E} \sqrt{1 - \frac{\xi_E^2 (2\epsilon_{\mathbf{w}_{up}} + \epsilon_{\mathbf{w}_{up}}^2)}{(1 + \epsilon_{\mathbf{w}_{up}})^2 g_E^2 \sigma^2}} - (1 + \epsilon_{\mathbf{w}_{up}}) \frac{g_E^2 \sigma^2}{\xi_E} + \epsilon_{\mathbf{w}_{up}} \xi_E$$

$$\approx (1 + \epsilon_{\mathbf{w}_{up}}) \frac{g_E^2 \sigma^2}{\xi_E} (1 - \frac{\xi_E^2 (2\epsilon_{\mathbf{w}_{up}} + \epsilon_{\mathbf{w}_{up}}^2)}{2(1 + \epsilon_{\mathbf{w}_{up}})^2 g_E^2 \sigma^2}) - (1 + \epsilon_{\mathbf{w}_{up}}) \frac{g_E^2 \sigma^2}{\xi_E} + \epsilon_{\mathbf{w}_{up}} \xi_E$$

$$= -\frac{\xi_E (2\epsilon_{\mathbf{w}_{up}} + \epsilon_{\mathbf{w}_{up}}^2)}{2(1 + \epsilon_{\mathbf{w}_{up}})} + \epsilon_{\mathbf{w}_{up}} \xi_E = \frac{2\epsilon_{\mathbf{w}_{up}} \xi_E (1 + \epsilon_{\mathbf{w}_{up}}) - \xi_E (2\epsilon_{\mathbf{w}_{up}} + \epsilon_{\mathbf{w}_{up}}^2)}{2(1 + \epsilon_{\mathbf{w}_{up}})} = \frac{\xi_E \epsilon_{\mathbf{w}_{up}}^2}{2(1 + \epsilon_{\mathbf{w}_{up}})}. \quad (121)$$

Combining (115), (120), and (121) we finally have

$$E\xi_{off}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, \mathbf{w}_{off}) - E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \geq \frac{\epsilon_{\mathbf{w}_{up}}^2}{2(1 + \epsilon_{\mathbf{w}_{up}})} E\xi_E \geq \frac{\epsilon_{\mathbf{w}_{up}}^2}{2(1 + \epsilon_{\mathbf{w}_{up}})} E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \quad (122)$$

where the last inequality follows by noting that in the definition of $\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ the elements of $\tilde{\mathbf{x}}$ and $\lambda^{(2)}$ are non-negative.

Now, roughly speaking, (122) shows that if $\|\mathbf{w}_{lasso}\|_2$ were to deviate from $\|\hat{\mathbf{w}}\|_2$ the optimal value of the objective in (6) would be higher than the lower bound derived in Section 2.1. We summarize these observations in the following lemma (essentially a deviating equivalent of Lemma 5 from Section 2.1).

Lemma 9. *Let $\mathbf{v}$ be an $n \times 1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m \times n$ matrix of i.i.d. standard normal random variables. Consider an $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x} = \tilde{\mathbf{x}}$. Let then ζ_{obj} be as defined in (10) or (14) and let $\mathbf{w}_{off}$ be the solution of (14). Let α and β be below the fundamental characterization (108) and let $\hat{\mathbf{w}}$ be as defined in (46). Assume that $\|\|\mathbf{w}_{off}\|_2 - \|\hat{\mathbf{w}}\|_2\| \geq \epsilon_{\mathbf{w}_{up}} \|\hat{\mathbf{w}}\|_2$, where $\epsilon_{\mathbf{w}_{up}}$ is an arbitrarily small but fixed constant. Then there would be a constant $\epsilon_{off} > 0$, and arbitrarily small positive constants $\epsilon_{lip}, \epsilon_1^{(\mathbf{h})}, \epsilon_1^{(g)}$ such that*

$$P(\zeta_{obj} \geq \zeta_{obj}^{(off)}) \geq 1 - e^{-\epsilon_{off} n}, \quad (123)$$

where

$$\zeta_{obj}^{(off)} = (1 - \epsilon_{lip})(1 + \frac{\epsilon_{\mathbf{w}_{up}}^2}{2(1 + \epsilon_{\mathbf{w}_{up}})}) E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) - \epsilon_1^{(\mathbf{h})} \sqrt{n} - \epsilon_1^{(g)} \sqrt{n}, \quad (124)$$

and $\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ is as defined in (42) (or (113)).

Proof. Follows from the previous discussion, discussion from Section 2.3.1, and a combination of (112), (115), (122), arguments right after (112), and Lemma 5. □

2.3.3 Deviation of the upper bound

In this section we will show that $\|\mathbf{w}_{lasso}\|_2$ can not deviate from $\|\hat{\mathbf{w}}\|_2$ as much as it was assumed in the previous section. To do so we will actually continue to assume that it can and then eventually reach a contradiction. As in previous section, let then $|\|\mathbf{w}_{off}\|_2 - \|\hat{\mathbf{w}}\|_2| \geq \epsilon_{\mathbf{w}_{up}}\|\hat{\mathbf{w}}\|_2$, where $\epsilon_{\mathbf{w}_{up}}$ is an arbitrarily small constant. Further, let $\xi_{dual}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ be

$$\xi_{dual}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{d \geq 0} \max_{\nu, \lambda^{(2)}} \quad \sqrt{d^2 + \sigma^2}\|\mathbf{g}\|_2 - d\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$
$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{125}$$

Rewriting (125) with a simple sign flipping turns out to be useful in what follows

$$-\xi_{dual}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{d \geq 0} \min_{\nu, \lambda^{(2)}} \quad -\sqrt{d^2 + \sigma^2}\|\mathbf{g}\|_2 + d\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$
$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{126}$$

The following lemma provides a powerful tool to deal with (126).

Lemma 10. *Let $\xi_{dual}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ be as defined in (126). Further, let*

$$-\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{\nu, \lambda^{(2)}} \max_{d \geq 0} \quad -\sqrt{d^2 + \sigma^2}\|\mathbf{g}\|_2 + d\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$
$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{127}$$

Then
$$\xi_{dual}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}). \tag{128}$$

Proof. After solving the inner maximization over d in (127) one has

$$d_{opt} = \sigma \frac{\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2}}. \tag{129}$$

Such a d then establishes that the right-hand side of (127) is indeed $\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$, i.e, one has as in (42)

$$-\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{\nu, \lambda^{(2)}} \quad -\sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$
$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{130}$$

Now we digress for a moment and consider the following optimization problem

$$\min_{\nu,\lambda^{(2)},\mathbf{q}_1,\mathbf{q}_2} \quad -\sigma\mathbf{q}_1 + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$
$$\text{subject to} \quad \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 \leq \mathbf{q}_2$$
$$\mathbf{q}_1^2 + \mathbf{q}_2^2 \leq \|\mathbf{g}\|_2^2$$
$$\nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{131}$$

Let $-\xi_{ov}^{(1)}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}})$ be the optimal value of its objective function. Let quadruplet $\hat{\nu},\widehat{\lambda^{(2)}},\hat{\mathbf{q}}_1,\hat{\mathbf{q}}_2$ be the solution of the above optimization problem. Then it must be

$$\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 = \hat{\mathbf{q}}_2 \tag{132}$$

and consequently

$$\hat{\mathbf{q}}_1 = \sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}$$
$$-\xi_{ov}^{(1)}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) = -\sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2} + \sum_{i=n-k+1}^{n} \widehat{\lambda_i^{(2)}}\tilde{\mathbf{x}}_i. \tag{133}$$

The above claim is rather obvious but for the completeness we sketch the argument that supports it. Assume that $\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 < \hat{\mathbf{q}}_2$, then $\hat{\mathbf{q}}_1 < \sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}$, and $-\xi_{ov}^{(1)}(\sigma,\mathbf{g},\mathbf{h})$ would be smaller then the expression on the right-hand side of (133). Now, since (132) and (133) hold one has that $-\xi_{ov}^{(1)}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}})$ can be determined through the following equivalent to (131)

$$-\xi_{ov}^{(1)}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) = \min_{\nu,\lambda^{(2)}} \quad -\sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$
$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n \tag{134}$$

After comparing (130) and (134) we have

$$-\xi_{ov}^{(1)}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) = -\xi_{ov}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}). \tag{135}$$

Now, let us write the Lagrange dual of the optimization problem in (131). Let d and γ_1 be Lagrangian variables such that

$$\max_{d \geq 0, \gamma_1 \geq 0} \min_{\nu,\lambda^{(2)},\mathbf{q}_1,\mathbf{q}_2} \quad -\sigma\mathbf{q}_1 + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i + d\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 - d\mathbf{q}_2 + \gamma_1(\mathbf{q}_1^2 + \mathbf{q}_2^2) - \gamma_1\|\mathbf{g}\|_2^2$$
$$\text{subject to} \quad \nu \geq 0$$
$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{136}$$

After solving the inner minimization over $\mathbf{q}_1, \mathbf{q}_2$ in (136) we have

$$\max_{d \geq 0, \gamma_1 \geq 0} \min_{\nu, \lambda^{(2)}} \quad -\frac{\sigma^2 + d^2}{4\gamma_1} - \gamma_1 \|\mathbf{g}\|_2^2 + \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i + d\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2$$

$$\text{subject to} \quad \nu \geq 0$$

$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{137}$$

Since the first two terms in the objective function in (137) do not involve neither ν nor $\lambda^{(2)}$ one can then maximize their sum over γ_1 for any d. After that we finally have

$$\max_{d \geq 0} \min_{\nu, \lambda^{(2)}} \quad -\sqrt{\sigma^2 + d^2}\|\mathbf{g}\|_2 + \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i + d\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2$$

$$\text{subject to} \quad \nu \geq 0$$

$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{138}$$

Let $-\xi_{ov}^{(2)}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ be the optimal value of the objective function in (138). Since (138) is the dual of (131) and since the strict duality obviously holds (the optimization problem in (131) is clearly convex) one has

$$-\xi_{ov}^{(2)}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = -\xi_{ov}^{(1)}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}). \tag{139}$$

On the other hand the optimization problem in (138) is the same as the one in (126) and therefore

$$-\xi_{ov}^{(2)}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = -\xi_{dual}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}). \tag{140}$$

Connecting (135), (139), and (140) one finally has

$$-\xi_{dual}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = -\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \tag{141}$$

which is what is stated in (128). This concludes the proof. $\qquad\square$

Let $\hat{d}, \hat{\nu}, \widehat{\lambda^{(2)}}$ be the solution of (125). Clearly, $\hat{d} = \|\hat{\mathbf{w}}\|_2 = \sigma \frac{\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}}$ and since all

quantities concentrate $E\hat{d} = E\|\hat{\mathbf{w}}\|_2 \doteq \sigma \frac{E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2}{\sqrt{E\|\mathbf{g}\|_2^2 - E\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}}$. Now, set $C_{\mathbf{w}_{up}} = E\|\hat{\mathbf{w}}\|_2$ in (90).
Then a combination of (90), (125), and Lemma 10 gives

$$E\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}, E\|\hat{\mathbf{w}}\|_2) = E \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0} (\sqrt{(E\|\hat{\mathbf{w}}\|_2)^2 + \sigma^2}\|\mathbf{g}\|_2 - E\|\hat{\mathbf{w}}\|_2\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i)$$

$$= E \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0} (\sqrt{(E\hat{d})^2 + \sigma^2}\|\mathbf{g}\|_2 - E\hat{d}\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i)$$

$$\doteq E \min_{d \geq 0} \max_{\lambda^{(2)} \in \Lambda^{(2)}, \nu \geq 0} (\sqrt{d^2 + \sigma^2}\|\mathbf{g}\|_2 - d\|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i) = E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}).$$

$$\tag{142}$$

Combining Lemma 14 and (142) one has that with overwhelming probability there is a $\mathbf{w}$ such that the objective in (6) is upper bounded by a quantity arbitrarily close from above to $E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$. On the

other hand Lemma 5 states that for any $\mathbf{w}$ such that $\|\|\mathbf{w}\|_2 - \|\hat{\mathbf{w}}\|_2\| \geq \epsilon_{w_{up}}\|\hat{\mathbf{w}}\|_2$, $\epsilon_{w_{up}} > 0$, the objective value of (6) is with overwhelming probability lower bounded by a quantity that is arbitrarily close from below to $(1 + \frac{\epsilon_{w_{up}}^2}{2(1+\epsilon_{w_{up}})})E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$. Clearly then the assumption of Lemma 5 is unsustainable and one has that $\|\mathbf{w}_{lasso}\|_2$ can not deviate substantially from $\|\hat{\mathbf{w}}\|_2$. This then implies that with overwhelming probability the objective value of (6) concentrates around $E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ and consequently that $\|\mathbf{w}_{lasso}\|_2$ concentrates around $E\|\hat{\mathbf{w}}\|_2$.

2.4 Connecting all pieces

In this section we connect all of the above. We will summarize the results obtained so far in the following theorem.

Theorem 1. *Let $\mathbf{v}$ be an $n \times 1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m \times n$ matrix of i.i.d. standard normal random variables. Further, let $\mathbf{g}$ and $\mathbf{h}$ be $m \times 1$ and $n \times 1$ vectors of i.i.d. standard normals, respectively. Consider a k-sparse $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x} = \tilde{\mathbf{x}}$. Let the solution of (6) be $\hat{\mathbf{x}}$ and let the so-called error vector of LASSO from (6) be $\mathbf{w}_{lasso} = \hat{\mathbf{x}} - \tilde{\mathbf{x}}$. Let n be large and let constants $\alpha = \frac{m}{n}$ and $\beta = \frac{k}{n}$ be below the fundamental characterization (108). Consider the following optimization problem:*

$$\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}} \quad \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$

$$\textit{subject to} \quad \nu \geq 0$$

$$0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n. \tag{143}$$

Let $\hat{\nu}$ and $\widehat{\lambda^{(2)}}$ be the solution of (143). Set

$$\|\hat{\mathbf{w}}\|_2 = \sigma\frac{\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}}. \tag{144}$$

Then:

$$P((1 - \epsilon_1^{(lasso)})E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \leq \|\mathbf{y} - A\hat{\mathbf{x}}\|_2 \leq (1 + \epsilon_1^{(lasso)})E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})) = 1 - e^{-\epsilon_2^{(lasso)}n} \tag{145}$$

and

$$P((1 - \epsilon_1^{(lasso)})E\|\hat{\mathbf{w}}\|_2 \leq \|\mathbf{w}_{lasso}\|_2 \leq (1 + \epsilon_1^{(lasso)})E\|\hat{\mathbf{w}}\|_2) = 1 - e^{-\epsilon_2^{(lasso)}n}, \tag{146}$$

where $\epsilon_1^{(lasso)} > 0$ is an arbitrarily small constant and $\epsilon_2^{(lasso)}$ is a constant dependent on $\epsilon_1^{(lasso)}$ and σ but independent of n.

Proof. Follows from the above discussion and a combination of (42), Lemma 2, discussion in Section 2.3.1, and Lemmas 5, 14, and 10. $\square$

It may not be clear immediately but the result presented in the above theorem is incredibly powerful. Among other things, it enables one to precisely estimate the norm of the error vector in "noisy" under-determined systems of linear equations. Moreover, it can do so for any given k-sparse vector $\tilde{\mathbf{x}}$. Furthermore, all of it is done through a transformation of the original LASSO from (6) to a much simpler optimization program (143). While many quantities of interest in LASSO recovery can be computed through the mechanism presented above, below we focus only on quantities that relate to what we will call LASSO's *generic*

performance. Computation of all other quantities that we consider are of interest will be presented in a series of forthcoming papers.

2.4.1 LASSO's generic performance

The results presented in the above theorem are fairly general and pertain to pretty much any possible scenario one can imagine. Here we will focus on the so-called "worst-case" scenario or as we will refer to it "generic" performance scenario. We will now show that $E\|\hat{\mathbf{w}}\|_2$ from Theorem 1 can be upper-bounded over the set of all $\tilde{\mathbf{x}}$'s. To that end let us assume that all nonzero components of $\tilde{\mathbf{x}}$ are infinite. The optimization problem from (143) then becomes

$$
\begin{aligned}
\xi_{ov}^{(gob)}(\sigma, \mathbf{g}, \mathbf{h}) = \max_{\nu, \lambda^{(2)}} \quad & \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} \\
\text{subject to} \quad & \nu \geq 0 \\
& 0 \leq \lambda_i^{(2)} = 0, n - k + 1 \leq i \leq n \\
& 0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n - k.
\end{aligned}
\tag{147}
$$

Let ν_{gen} and $\lambda^{(gen)}$ be the solution of (147) and let $\mathbf{w}_{gen}$ be the error vector in case when all nonzero compoenents of $\tilde{\mathbf{x}}$ are infinite (in Section 2.3.1 for a slightly changed version of (147) ν_{gen} and $\lambda^{(gen)}$ were referred to as ν_{ℓ_1} and $\lambda^{(\ell_1)}$). Now, let us assume that some of nonzero components of $\tilde{\mathbf{x}}$ in (143) are finite. And let as usual $\hat{\nu}$ and $\widehat{\lambda^{(2)}}$ be the solution of (143) and let $\|\hat{\mathbf{w}}\|_2$ be the norm of the LASSO's error vector. Since

$$
\sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2} - \sum_{i=n-k+1}^{n} \widehat{\lambda_i^{(2)}}\tilde{\mathbf{x}}_i \geq \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu_{gen}\mathbf{z}^{(1)} - \lambda^{(gen)}\|_2^2}
\tag{148}
$$

and $\widehat{\lambda_i^{(2)}} \geq 0, n - k + 1 \leq i \leq n$, one has that

$$
\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2 \leq \|\mathbf{h} + \nu_{gen}\mathbf{z}^{(1)} - \lambda^{(gen)}\|_2.
\tag{149}
$$

Furthermore, one then has for the norm of error vectors

$$
\|\hat{\mathbf{w}}\|_2 = \sigma\frac{\|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \hat{\nu}\mathbf{z}^{(1)} - \widehat{\lambda^{(2)}}\|_2^2}} \leq \sigma\frac{\|\mathbf{h} + \nu_{gen}\mathbf{z}^{(1)} - \lambda^{(gen)}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu_{gen}\mathbf{z}^{(1)} - \lambda^{(gen)}\|_2^2}} = \|\mathbf{w}_{gen}\|_2.
\tag{150}
$$

Then the following *generic* equivalent to Theorem 1 can be established.

Theorem 2. *Assume the setup of Theorem 1. Consider the following optimization problem:*

$$
\begin{aligned}
\xi_{ov}^{(gen)}(\sigma, \mathbf{g}, \mathbf{h}) = \min_{\nu, \lambda^{(2)}} \quad & \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2 \\
\textit{subject to} \quad & \nu \geq 0 \\
& \lambda_i^{(2)} = 0, n - k + 1 \leq i \leq n \\
& 0 \leq \lambda_i^{(2)} \leq 2\nu, 1 \leq i \leq n - k.
\end{aligned}
\tag{151}
$$

Let ν_{gen} and $\lambda^{(gen)}$ be the solution of (151). Set

$$\|\mathbf{w}_{gen}\|_2 = \sigma \frac{\|\mathbf{h} + \nu_{gen}\mathbf{z}^{(1)} - \lambda^{(gen)}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu_{gen}\mathbf{z}^{(1)} - \lambda^{(gen)}\|_2^2}}. \tag{152}$$

Then:

$$
\begin{aligned}
P(\exists \mathbf{w}_{lasso} | \|\mathbf{w}_{lasso}\|_2 \in ((1 - \epsilon_1^{(lasso)})E\|\mathbf{w}_{gen}\|_2, (1 + \epsilon_1^{(lasso)})E\|\mathbf{w}_{gen}\|_2)) &\geq 1 - e^{-\epsilon_2^{(lasso)}n} \\
P(\|\mathbf{w}_{lasso}\|_2 \leq (1 + \epsilon_1^{(lasso)})E\|\mathbf{w}_{gen}\|_2)) &\geq 1 - e^{-\epsilon_3^{(lasso)}n},
\end{aligned}
\tag{153}
$$

where $\epsilon_1^{(lasso)} > 0$ is an arbitrarily small constant and $\epsilon_2^{(lasso)}$ and $\epsilon_3^{(lasso)}$ are constants dependent on $\epsilon_1^{(lasso)}$ and σ but independent of n.

Proof. Follows from the above discussion, Theorem 1, and by noting that the optimization problems in (151) and (147) are equivalent. $\square$

The following corollary then provides a quick way of computing the concentrating point of the "worst case" norm of the error vector.

Corollary 1. *Assume the setup of Theorems 1 and 2. Let $\alpha = \frac{m}{n}$ and $\beta_w = \frac{k}{n}$. Then*

$$
\begin{aligned}
P(\exists \mathbf{w}_{lasso} | \|\mathbf{w}_{lasso}\|_2 \in ((1 - \epsilon_1^{(lasso)})\sigma\sqrt{\frac{\alpha_w}{\alpha - \alpha_w}}, (1 + \epsilon_1^{(lasso)})\sigma\sqrt{\frac{\alpha_w}{\alpha - \alpha_w}})) &\geq 1 - e^{-\epsilon_2^{(lasso)}n} \\
P(\|\mathbf{w}_{lasso}\|_2 \leq (1 + \epsilon_1^{(lasso)})\sigma\sqrt{\frac{\alpha_w}{\alpha - \alpha_w}}) &\geq 1 - e^{-\epsilon_2^{(lasso)}n},
\end{aligned}
\tag{154}
$$

where $\alpha_w < \alpha$ is such that

$$(1 - \beta_w)\frac{\sqrt{\frac{2}{\pi}}e^{-(\text{erfinv}(\frac{1 - \alpha_w}{1 - \beta_w}))^2}}{\alpha_w} - \sqrt{2}\text{erfinv}(\frac{1 - \alpha_w}{1 - \beta_w}) = 0 \tag{155}$$

and $\epsilon_1^{(lasso)} > 0$ is an arbitrarily small constant and $\epsilon_2^{(lasso)}$ is a constant dependent on $\epsilon_1^{(lasso)}$ and σ but independent of n.

Proof. Let $\bar{\mathbf{h}}$ and $\mathbf{z}^{(2)}$ be as in Section 2.3.1. Then

$$
\begin{aligned}
\xi_{ov}^{(gen)}(\sigma, \mathbf{g}, \mathbf{h}) = \min_{\nu, \lambda^{(2)}} \quad & \|\bar{\mathbf{h}} - \nu\mathbf{z}^{(2)} + \lambda^{(2)}\|_2 \\
\text{subject to} \quad & \nu \geq 0 \\
& \lambda_i^{(2)} = 0, n - k + 1 \leq i \leq n \\
& 0 \leq \lambda_i^{(2)} \leq \nu, 1 \leq i \leq n - k,
\end{aligned}
\tag{156}
$$

is equivalent to (154). Moreover

$$
\begin{aligned}
E\|\mathbf{w}_{gen}\|_2 &\doteq \sigma \frac{E\xi_{ov}^{(gen)}(\sigma, \mathbf{g}, \mathbf{h})}{\sqrt{E\|\mathbf{g}\|_2^2 - E\xi_{ov}^{(gen)}(\sigma, \mathbf{g}, \mathbf{h})^2}} \\
&\doteq \sigma\sqrt{\frac{\alpha_w}{\alpha - \alpha_w}},
\end{aligned}
\tag{157}
$$

where $\alpha_w m \doteq E\xi_{ov}^{(gen)}(\sigma, \mathbf{g}, \mathbf{h})^2$ is one of the main contributions of [63]. The rest then trivially follows from (153). $\square$

Using (155) and (147) one can then for any σ and any pair (α, β_w) (that is below fundamental characterization (108)) determine the value of the worst case $E\|\mathbf{w}_{lasso}\|_2$ as $\sigma\sqrt{\frac{\alpha_w}{\alpha - \alpha_w}}$. We present the obtained results in Figure 3. For several fixed values of worst case $E\|\mathbf{w}_{lasso}\|_2$ we determine curves of points (α, β_w) for which these fixed values are achieved (of course for any α that is below a curve the value of the corresponding worst case $E\|\mathbf{w}_{lasso}\|_2$ is smaller). As can be seen from the plots the lower the norm-2 of the error vector the smaller the allowable region for pairs (α, β_w).

The results of the above corollary match those obtained in [7, 26] through a state evolution/bilief propagation type of analysis. The above corollary relates to the LASSO from (6) whereas the results from [7, 26] are derived for somewhat different LASSO from (5). However, as mentioned earlier, in Section 4 we will establish a nice connection between the LASSO from (6) and one that is fairly similar to (5).
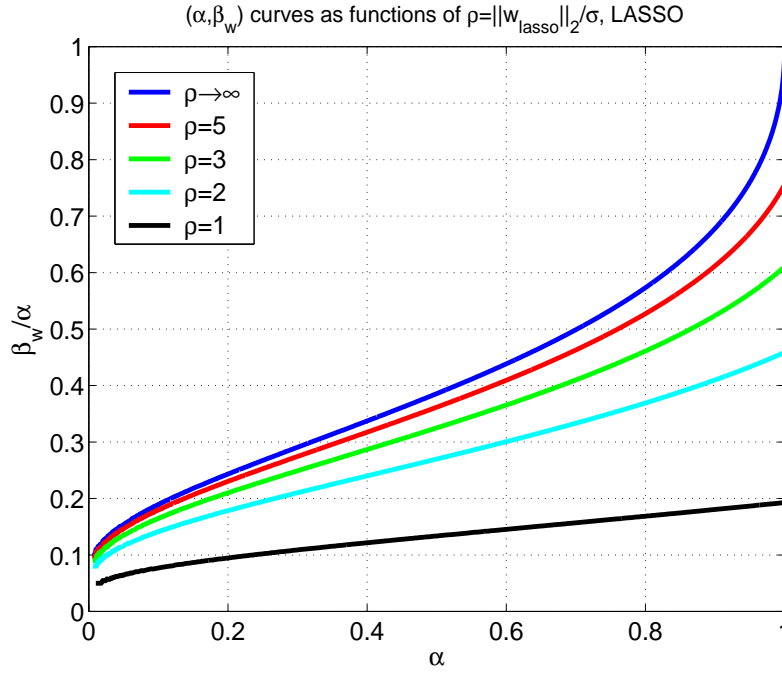

Figure 3: (α, β_w) curves as functions of $\rho = \frac{E\|\mathbf{w}_{lasso}\|_2}{\sigma}$ for LASSO algorithm from (6)

34

3   LASSO's performance analysis framework – signed $\mathbf{x}$

In this section we show how the LASSO's performance analysis framework developed in the previous section can be specialized to the case when signals are *a priori* known to have nonzero components of certain sign. All major assumptions stated at the beginning of the previous section will continue to hold in this section as well; namely, we will continue to consider matrices A with i.i.d. standard normal random variables; elements of $\mathbf{v}$ will again be i.i.d. Gaussian random variables with zero mean and variance σ. The main difference, though, comes in the definition of $\tilde{\mathbf{x}}$. We will in this section assume that $\tilde{\mathbf{x}}$ is the original $\mathbf{x}$ in (3) that we are trying to recover and that it is *any* k-sparse vector with a given fixed location of its nonzero elements and with a priori known signs of its elements. Given the statistical context, it will be fairly easy to see later on that everything that we will present in this section will be irrelevant with respect to what particular location and what particular combination of signs of nonzero elements are chosen. We therefore for the simplicity of the exposition and without loss of generality assume that the components $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-k}$ of $\mathbf{x}$ are equal to zero and the components $\mathbf{x}_{n-k+1}, \mathbf{x}_{n-k+2}, \ldots, \mathbf{x}_n$ of $\mathbf{x}$ are greater than or equal to zero. However, differently from what was assumed in the previous section, we now assume that this information is *a priori* known. That essentially means that this information is also known to the solving algorithm. Then instead of (6) one can consider its a better ("signed") version

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & \|\mathbf{y} - A\mathbf{x}\|_2 \\
\text{subject to} \quad & \|\mathbf{x}\|_1 \leq \|\tilde{\mathbf{x}}\|_1 \\
& \mathbf{x}_i \geq 0, 1 \leq i \leq n.
\end{aligned} \tag{158}
$$

In what follows we will mimic the procedure presented in the previous section, skip all the obvious parallels, and emphasize the points that are different. The framework that we will present below will again center around finding the optimal value of the objective function in (158). In the first of the following two subsections we will create a lower bound on this optimal value (this will essentially amount to creating a procedure that is analogous to the one presented in Section 2.1). We will then afterwards in the second of the subsections create an upper bound on this optimal value. As it was done in the case of general $\tilde{\mathbf{x}}$ in the previous section we will in the third subsection show that the two bounds actually match. To make further writing easier and clearer we set already here

$$
\begin{aligned}
\zeta_{obj+} = \min_{\mathbf{x}} \quad & \|\mathbf{y} - A\mathbf{x}\|_2 \\
\text{subject to} \quad & \|\mathbf{x}\|_1 \leq \|\tilde{\mathbf{x}}\|_1 \\
& \mathbf{x}_i \geq 0, 1 \leq i \leq n.
\end{aligned} \tag{159}
$$

3.1   Lower-bounding ζ_{obj+}

In this section we present the part of the framework that relates to finding a "high-probability" lower bound on ζ_{obj}^+. As in the previous section we again assume that there is a (if necessary, arbitrarily large) constant $C_\mathbf{w}$ such that

$$
P(\|\mathbf{w}_{lasso}\|_2 \leq C_\mathbf{w}) = 1 - e^{-\epsilon_{C_\mathbf{w}} n}. \tag{160}
$$

We again start by noting that if one knows that $\mathbf{y} = A\tilde{\mathbf{x}} + \mathbf{v}$ holds then (159) can be rewritten as

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & \|\mathbf{v} + A\tilde{\mathbf{x}} - A\mathbf{x}\|_2 \\
\text{subject to} \quad & \|\mathbf{x}\|_1 \leq \|\tilde{\mathbf{x}}\|_1 \\
& \mathbf{x}_i \geq 0, 1 \leq i \leq n.
\end{aligned} \tag{161}
$$

After a small change of variables, $\mathbf{x} = \tilde{\mathbf{x}} + \mathbf{w}$, one has an equivalent to (14)

$$\min_{\mathbf{w}} \quad \|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \|_2$$

$$\text{subject to} \quad \sum_{i=1}^{n} \mathbf{w}_i \leq 0$$

$$\tilde{\mathbf{x}}_i + \mathbf{w}_i \geq 0, 1 \leq i \leq n, \tag{162}$$

where as earlier $A_{\mathbf{v}} = \begin{bmatrix} -A & \mathbf{v} \end{bmatrix}$ is an $m \times (n+1)$ random matrix with i.i.d. standard normal components. Let

$$S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}}) = \{ \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \in R^{n+1} | \quad \|\mathbf{w}\|_2 \leq C_{\mathbf{w}} \quad \text{and} \quad \sum_{i=1}^{n} \mathbf{w}_i \leq 0 \quad \text{and} \quad \tilde{\mathbf{x}}_i + \mathbf{w}_i \geq 0, 1 \leq i \leq n \}. \tag{163}$$

Further, let

$$f_{obj+}(\sigma, \mathbf{w}) = \|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \|_2 \tag{164}$$

and set,

$$\zeta_{obj+}^{(help)} = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} f_{obj+}(\sigma, \mathbf{w}) = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \|_2 = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \max_{\|\mathbf{a}\|_2=1} \mathbf{a}^T A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix}. \tag{165}$$

Now, after applying Lemma 1 and following the procedure from the previous section one has

$$P(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} (f_{obj+}(\sigma, \mathbf{w}) + \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} g) \geq \zeta_{obj+}^{(l)}) \geq p_l^+. \tag{166}$$

where

$$p_l^+ = P \left(\min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \|\mathbf{g}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i \right) + \mathbf{h}_{n+1} \sigma \geq \zeta_{obj+}^{(l)} \right). \tag{167}$$

As in previous section we will essentially show that for certain $\zeta_{obj+}^{(l)}$ this probability is close to 1 which will imply that we have a "high probability" lower bound on ζ_{obj+}. Let

$$\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{[\mathbf{w}^T \sigma]^T \in S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})} \left(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \|\mathbf{g}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i \right). \tag{168}$$

Now we split the analysis into two parts. The first one will be the deterministic analysis of $\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ and will be presented in Subsection 3.1.1. In the second part (that will be presented in Subsection 3.1.2) we will use the results of such a deterministic analysis and continue the above probabilistic analysis applying various concentration results.

3.1.1 Optimizing $\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$

In this section we compute $\xi_+(\sigma, \mathbf{g}, \mathbf{h})$. We first rewrite the optimization problem from (168) in the following way

$$
\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{\mathbf{w}} \quad \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \|\mathbf{g}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i
$$

$$
\text{subject to} \quad \sum_{i=1}^{n} \mathbf{w}_i \leq 0
$$

$$
\tilde{\mathbf{x}}_i + \mathbf{w}_i \geq 0, 1 \leq i \leq n
$$

$$
\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \leq \sqrt{C_{\mathbf{w}}^2 + \sigma^2}. \tag{169}
$$

The Lagrange dual of the above problem then becomes

$$
\mathcal{L}(\nu, \lambda^{(2)}, \mathbf{w}, \gamma) = \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \|\mathbf{g}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i + \nu \sum_{i=1}^{n} \mathbf{w}_i
$$

$$
- \sum_{i=n-k+1}^{n} \lambda_i^{(2)} (\tilde{\mathbf{x}}_i + \mathbf{w}_i) - \sum_{i=1}^{n-k} \lambda_i^{(2)} \mathbf{w}_i + \gamma(\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} - \sqrt{C_{\mathbf{w}}^2 + \sigma^2}). \tag{170}
$$

After a few further arrangements we finally have

$$
\mathcal{L}(\nu, \lambda^{(2)}, \mathbf{w}, \gamma) = \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}(\|\mathbf{g}\|_2 + \gamma) + \sum_{i=1}^{n}(\mathbf{h}_i + \nu - \lambda_i^{(2)})\mathbf{w}_i - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i - \gamma\sqrt{C_{\mathbf{w}}^2 + \sigma^2}. \tag{171}
$$

One can then write the following dual problem of (169)

$$
\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}, \gamma} \min_{\mathbf{w}} \quad \mathcal{L}(\nu, \lambda^{(2)}, \mathbf{w})
$$

$$
\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n
$$

$$
\nu \geq 0
$$

$$
\gamma \geq 0, \tag{172}
$$

where we of course use the fact that the strict duality obviously holds. The inner minimization over $\mathbf{w}$ is now doable. Setting the derivatives with respect to $\mathbf{w}_i$ to zero one obtains

$$
\frac{\mathbf{w}(\|\mathbf{g}\|_2 + \gamma)}{\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}} + (\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}) = 0, \tag{173}
$$

where $\mathbf{z}^{(1)}$ and $\lambda^{(2)}$ are as defined in the previous section. From (173) one then has

$$
\mathbf{w}(\|\mathbf{g}\|_2 + \gamma) = -\sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}(\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}) \tag{174}
$$

or in a norm form

$$
\|\mathbf{w}\|_2^2(\|\mathbf{g}\|_2 + \gamma)^2 = (\|\mathbf{w}\|_2^2 + \sigma^2)\|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2. \tag{175}
$$

From (175) we then find

$$\|\mathbf{w}_{sol+}\|_2 = \frac{\sigma \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2}{\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2}}, \tag{176}$$

and from (174)

$$\mathbf{w}_{sol+} = \frac{\sigma(\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)})}{\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2}} \tag{177}$$

where $\mathbf{w}_{sol+}$ is of course the solution of the inner minimization over $\mathbf{w}$. As in the previous section, one should note that (176) and (177) are of course possible only if $\|\mathbf{g}\|_2 + \gamma - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2 \geq 0$. (Also, as in the previous section if for ν and $\lambda^{(2)}$ that are optimal in (172) the condition is not met then for the corresponding (α, β) the worst-case $\|\mathbf{w}\|_2$ is infinite with overwhelming probability). Plugging the value of $\mathbf{w}_{sol+}$ from (177) back in (172) gives

$$\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}, \gamma} \quad \sigma \sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i - \gamma \sqrt{C_{\mathbf{w}}^2 + \sigma^2}$$

$$\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n$$
$$\nu \geq 0$$
$$\|\mathbf{g}\|_2 + \gamma - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2 \geq 0$$
$$\gamma \geq 0. \tag{178}$$

Now, the maximization over γ can be done. After setting the derivative to zero one finds

$$\frac{\|\mathbf{g}\|_2 + \gamma}{\sqrt{(\|\mathbf{g}\|_2 + \gamma)^2 - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2}} - \sqrt{C_{\mathbf{w}}^2 + \sigma^2} = 0 \tag{179}$$

and after some algebra

$$\gamma_{opt+} = \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}} \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2 - \|\mathbf{g}\|_2, \tag{180}$$

where of course γ_{opt+} would be the solution of (178) only if larger than or equal to zero. Alternatively of course $\gamma_{opt+} = 0$. Now, based on these two scenarios we distinguish two different optimization problems:

1. *The "overwhelming" optimization — signed $\tilde{\mathbf{x}}$*

$$\xi_{ov+}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}} \quad \sigma \sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \nu \geq 0$$
$$\lambda_i^{(2)} \geq 0, 1 \leq i \leq n. \tag{181}$$

2. *The "non-overwhelming" optimization — signed $\tilde{\mathbf{x}}$*

$$\xi_{nov+}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}} \quad \sqrt{C_{\mathbf{w}}^2 + \sigma^2} \|\mathbf{g}\|_2 - C_{\mathbf{w}} \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \nu \geq 0$$
$$\lambda_i^{(2)} \geq 0, 1 \leq i \leq n. \tag{182}$$

The "overwhelming" optimization is the equivalent to (178) if for its optimal values $\widehat{\nu^+}$ and $\widehat{\lambda^{(2+)}}$ holds

$$\sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}} \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2 \leq \|\mathbf{g}\|_2, \tag{183}$$

We now summarize in the following lemma the results of this subsection.

Lemma 11. *Let $\widehat{\nu^+}$ and $\widehat{\lambda^{(2+)}}$ be the solutions of (181) and analogously let $\widetilde{\nu^+}$ and $\widetilde{\lambda^{(2+)}}$ be the solutions of (182). Let $\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ be, as defined in (168), the optimal value of the objective function in (168). Then*

$$\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \begin{cases} \sigma \sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2^2} - \sum_{i=n-k+1}^{n} \widehat{\lambda_i^{(2+)}} \tilde{\mathbf{x}}_i, & \text{if } \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}} \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2 \leq \|\mathbf{g}\|_2 \\ \sqrt{C_{\mathbf{w}}^2 + \sigma^2} \|\mathbf{g}\|_2 - C_{\mathbf{w}} \|\mathbf{h} + \widetilde{\nu^+} \mathbf{z}^{(1)} - \widetilde{\lambda^{(2+)}}\|_2 - \sum_{i=n-k+1}^{n} \widetilde{\lambda_i^{(2+)}} \tilde{\mathbf{x}}_i, & \text{otherwise} \end{cases}.$$
$$\tag{184}$$

Moreover, let $\widehat{\mathbf{w}^+}$ be the solution of (168). Then

$$\widehat{\mathbf{w}^+}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \begin{cases} \dfrac{\sigma(\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}})}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2^2}}, & \text{if } \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}} \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2 \leq \|\mathbf{g}\|_2 \\ \dfrac{C_{\mathbf{w}}(\mathbf{h} + \widetilde{\nu^+} \mathbf{z}^{(1)} - \widetilde{\lambda^{(2+)}})}{\|\mathbf{h} + \widetilde{\nu^+} \mathbf{z}^{(1)} - \widetilde{\lambda^{(2+)}}\|_2}, & \text{otherwise} \end{cases}, \tag{185}$$

and

$$\|\widehat{\mathbf{w}^+}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})\|_2 = \begin{cases} \dfrac{\sigma \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}})\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2^2}}, & \text{if } \sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}} \|\mathbf{h} + \widehat{\nu^+} \mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2 \leq \|\mathbf{g}\|_2 \\ C_{\mathbf{w}}, & \text{otherwise} \end{cases}.$$
$$\tag{186}$$

Proof. The first part follows trivially. The second one follows from (177) by choosing the optimal $\widehat{\nu^+}$ and $\widehat{\lambda^{(2+)}}$ or alternatively $\widetilde{\nu^+}$ and $\widetilde{\lambda^{(2+)}}$. $\square$

3.1.2 Concentration of $\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$

In this section we establish that $\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ concentrates with high probability around its mean. The following lemma is an analogue to Lemma 3.1.2

Lemma 12. *Let $\mathbf{g}$ and $\mathbf{h}$ be m and n dimensional vectors, respectively, with i.i.d. standard normal variables as their components. Let $\sigma > 0$ be an arbitrary scalar. Let $\xi(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ be as in (22). Further let $\epsilon_{lip} > 0$ be any constant. Then*

$$P(|\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) - E\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})| \geq \epsilon_{lip} E\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})) \leq \exp\left\{-\frac{(\epsilon_{lip} E\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}))^2}{2(2C_{\mathbf{w}}^2 + \sigma^2)}\right\}. \tag{187}$$

Proof. It follows by literally repeating every step of proof of Lemma . The only difference is that one now has $S_{\mathbf{w}}^+(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})$ instead of $S_{\mathbf{w}}(\sigma, \tilde{\mathbf{x}}, C_{\mathbf{w}})$. $\qquad\square$

Moreover one then has that $\|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2$ and $\|\mathbf{h} + \widetilde{\nu^+}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2+)}}\|_2$ concentrate as well which automatically implies that $\widehat{\mathbf{w}^+}$ also concentrates. More formally, one then has analogues to (187)

$$
\begin{aligned}
P(|\|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2 - E\|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2| \geq \epsilon_1^{(norm)} E\|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2) &\leq e^{-\epsilon_2^{(norm)}n} \\
P(|\|\mathbf{h} + \widetilde{\nu^+}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2+)}}\|_2 - E\|\mathbf{h} + \widetilde{\nu^+}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2+)}}\|_2| \geq \epsilon_3^{(norm)} E\|\mathbf{h} + \widetilde{\nu^+}\mathbf{z}^{(1)} - \widetilde{\lambda^{(2+)}}\|_2) &\leq e^{-\epsilon_4^{(norm)}n} \\
P(|\|\widehat{\mathbf{w}^+}\|_2 - E\|\widehat{\mathbf{w}^+}\|_2| \geq \epsilon_1^{(\mathbf{w})} E\|\widehat{\mathbf{w}^+}\|_2) &\leq e^{-\epsilon_2^{(\mathbf{w})}n},
\end{aligned}
\tag{188}
$$

where as usual $\epsilon_1^{(norm)} > 0$, $\epsilon_2^{(norm)} > 0$, and $\epsilon_1^{(\mathbf{w})} > 0$ are arbitrarily small constants and $\epsilon_3^{(norm)}$, $\epsilon_4^{(norm)}$, and $\epsilon_2^{(\mathbf{w})}$ are constant dependent on $\epsilon_1^{(norm)} > 0$, $\epsilon_2^{(norm)} > 0$, and $\epsilon_1^{(\mathbf{w})} > 0$, respectively, but independent of n. After repeating every step between (55) and (64) one arrives to the following analogue to Lemma 5.

Lemma 13. *Let $\mathbf{v}$ be an $n \times 1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m \times n$ matrix of i.i.d. standard normal random variables. Consider an $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x} = \tilde{\mathbf{x}}$. Let then ζ_{obj+} be as defined in (159) and let $\mathbf{w}^+$ be the solution of (162). Assume $P(\|\mathbf{w}^+\|_2 \leq C_{\mathbf{w}}) \geq 1 - e^{-\epsilon_{C_{\mathbf{w}}}n}$ for an arbitrarily large constant $C_{\mathbf{w}}$ and a constant $\epsilon_{C_{\mathbf{w}}} > 0$ dependent on $C_{\mathbf{w}}$ but independent of n. Then there is a constant $\epsilon_{lower} > 0$*

$$
P(\zeta_{obj+} \geq \zeta_{obj+}^{(lower)}) \geq (1 - e^{-\epsilon_{lower}n})(1 - e^{-\epsilon_{C_{\mathbf{w}}}n}),
\tag{189}
$$

where

$$
\zeta_{obj+}^{(lower)} = (1 - \epsilon_{lip})E\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) - \epsilon_1^{(\mathbf{h})}\sqrt{n} - \epsilon_1^{(g)}\sqrt{n},
\tag{190}
$$

$\xi_+(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})$ *is as defined in (168), and $\epsilon_{lip}, \epsilon_1^{(\mathbf{h})}, \epsilon_1^{(g)}$ are all positive arbitrarily small constants.*

Proof. Follows from the previous discussion. $\qquad\square$

3.2 Upper-bounding ζ_{obj+}

In this section we present a general framework for finding a "high-probability" upper bound on ζ_{obj+}. To that end, let r_+ and $C_{\mathbf{w}_{up+}}$ be positive scalars (as in Section 3.2, we in this subsection present a general framework and take these scalars to be arbitrary; however to make the bound as tight sa possible in the following subsection we will make them take particular values). As earlier, if we can show that there is a $\mathbf{w} \in R^n$ such that $\|\tilde{\mathbf{x}} + \mathbf{w}\|_1 \leq \|\tilde{\mathbf{x}}\|_1$ and $\|\mathbf{v} - A\mathbf{w}\|_2 \leq r_+$ with overwhelming probability then r_+ can act as an upper bound on ζ_{obj+}. We then start by looking at the following optimization problem

$$
\begin{aligned}
\min_{\mathbf{w}} \quad & \|\tilde{\mathbf{x}} + \mathbf{w}\|_1 - \|\tilde{\mathbf{x}}\|_1 \\
& \left\|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix}\right\|_2 \leq r_+ \\
& \tilde{\mathbf{x}}_i + \mathbf{w}_i \geq 0, 1 \leq i \leq n \\
& \|\mathbf{w}\|_2^2 \leq C_{\mathbf{w}_{up+}}^2,
\end{aligned}
\tag{191}
$$

where $A_{\mathbf{v}}$ is as defined right after (14). If we can show that with overwhelming probability the objective value of the above optimization problem is negative then r_+ will be a valid "high probability" upper-bound on ζ_{obj+}. Moreover, it will be achieved by a $\mathbf{w}$ for which it will hold that $\|\mathbf{w}\|_2 \leq C_{\mathbf{w}_{up+}}$.

First let us rewrite the objective value of the above optimization problem in a slightly more convenient form

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & \sum_{i=1}^{n} \mathbf{w}_i \\
& \left\| A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \right\|_2 \leq r_+ \\
& \tilde{\mathbf{x}}_i + \mathbf{w}_i \geq 0, 1 \leq i \leq n \\
& \|\mathbf{w}\|_2^2 \leq C_{\mathbf{w}_{up+}}^2.
\end{aligned} \tag{192}
$$

Now, we proceed in a fashion similar to the one from Subsection 3.1.1. We first do a slight modification of the first constraint from (192) in the following way

$$
\begin{aligned}
\min_{\mathbf{x}} \quad & \sum_{i=1}^{n} \mathbf{w}_i \\
& \left\| A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \right\|_2 \leq r \\
& \tilde{\mathbf{x}}_i + \mathbf{w}_i \geq 0, 1 \leq i \leq n \\
& \|\mathbf{b}\|_2^2 \leq r_+^2 \\
& \begin{bmatrix} -A\mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} = b \\
& \|\mathbf{w}\|_2^2 \leq C_{\mathbf{w}_{up+}}^2.
\end{aligned} \tag{193}
$$

The Lagrange dual of the above problem then becomes

$$
\mathcal{L}(\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{b}) = \sum_{i=1}^{n} \mathbf{w}_i - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}(\tilde{\mathbf{x}}_i + \mathbf{w}_i)
$$
$$
- \sum_{i=1}^{n-k} \lambda_i^{(2)} \mathbf{w}_i - \nu^{(1)} A\mathbf{w} + \nu^{(1)} \mathbf{v}\sigma - \nu^{(1)} \mathbf{b} + \gamma_1 (\sum_{i=1}^{n} \mathbf{b}_1^2 - r_+^2) + \gamma_2 (\|\mathbf{w}\|_2^2 - C_{\mathbf{w}_{up+}}^2), \quad (194)
$$

where $\nu^{(1)}$ and $\lambda^{(2)}$ are vectors of Lagrange variables as in previous sections. After rearranging terms we further have

$$
\mathcal{L}(\lambda^{(2)}, \nu^{(1)}, \gamma_1, \gamma_2, \mathbf{w}, \mathbf{b}) = - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i
$$
$$
+ ((\mathbf{z}^{(1)} - \lambda^{(2)})^T - \nu^{(1)} A)\mathbf{w} + \nu^{(1)} \mathbf{v}\sigma - \nu^{(1)} \mathbf{b} + \gamma_1 (\sum_{i=1}^{n} \mathbf{b}_1^2 - r_+^2) + \gamma_2 (\|\mathbf{w}\|_2^2 - C_{\mathbf{w}_{up+}}^2). \quad (195)
$$

Finally we can write a dual problem to (193)

$$\max_{\lambda^{(2)},\nu^{(1)},\gamma_1,\gamma_2} \min_{\mathbf{w},\mathbf{b}} \quad \mathcal{L}(\lambda^{(2)},\nu^{(1)},\gamma_1,\gamma_2,\mathbf{w},\mathbf{b})$$

$$\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n$$
$$\gamma_1 \geq 0,$$
$$\gamma_2 \geq 0, \tag{196}$$

where we of course use the fact that the strict duality obviously holds. Now, after repeating all the steps from (75) to (84) (wherever we had $2\lambda^{(2)}$ we would now have $\lambda^{(2)}$ and there will be no upper bound on components of $\lambda^{(2)}$ in the corresponding optimization problems) one obtains and analogue to (84)

$$-\min_{\lambda^{(2)},\nu^{(1)}} \max_{\|\mathbf{a}\|_2 = C_{\mathbf{w}_{up+}}} \quad ((\mathbf{z}^{(1)} - \lambda^{(2)})^T - \nu^{(1)}A)\mathbf{a} - \nu^{(1)}\mathbf{v}\sigma + \|\nu^{(1)}\|_2 r_+ + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n. \tag{197}$$

Now let us define $f_{obj+}^{(up)}$ as

$$-f_{obj+}^{(up)} = -\min_{\lambda^{(2)},\nu^{(1)}} \max_{\|\mathbf{a}\|_2 = C_{\mathbf{w}_{up+}}} \quad ((\mathbf{z}^{(1)} - \lambda^{(2)})^T - \nu^{(1)}A)\mathbf{a} - \nu^{(1)}\mathbf{v}\sigma + \|\nu^{(1)}\|_2 r_+ + \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n. \tag{198}$$

Any r_+ such that $\lim_{n\to} P(f_{obj+}^{(up)} \geq 0) = 1$ is then a valid "high-probability" upper bound. Set

$$\Lambda^{(2+)} = \{\lambda^{(2)} \in R^n | \lambda_i^{(2)} \geq 0, 1 \leq i \leq n\}, \tag{199}$$

and

$$\xi_{up+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},C_{\mathbf{w}_{up}}) = \max_{\lambda^{(2)}\in\Lambda^{(2+)},\nu\geq 0}(\sqrt{C_{\mathbf{w}_{up+}}^2 + \sigma^2}\|\mathbf{g}\|_2 - C_{\mathbf{w}_{up+}}\|\mathbf{h}+\nu\mathbf{z}^{(1)}-\lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i).$$
$$\tag{200}$$

After further repeating all the steps between (84) and Lemma 14 (the only difference is that $\lambda_i^{(2)} \in \Lambda^{(2+)}$ in the "signed" scenario) one then has the following "signed" analogue to Lemma 14 (which in essence gives a way of finding an r_+ such that $\lim_{n\to} P(f_{obj+}^{(up)} \geq 0) = 1$).

Lemma 14. *Let $\mathbf{v}$ be an $n \times 1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m \times n$ matrix of i.i.d. standard normal random variables. Consider an $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x} = \tilde{\mathbf{x}}$. Let then ζ_{obj+} be as defined in (159) and let $\mathbf{w}^+$ be the solution of (162). There is a constant $\epsilon_{upper} > 0$*

$$P(\zeta_{obj+} \leq \zeta_{obj+}^{(upper)}) \geq 1 - e^{-\epsilon_{upper}n}, \tag{201}$$

where

$$\zeta_{obj+}^{(upper)} = (1 + \epsilon_{lip})E\xi_{up+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},C_{\mathbf{w}_{up+}}) + \epsilon_1^{(\mathbf{h})}\sqrt{n} + \epsilon_3^{(g)}\sqrt{n}, \tag{202}$$

$\xi_{up+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},C_{\mathbf{w}_{up+}})$ is as defined in (200), $\epsilon_{lip}, \epsilon_1^{(\mathbf{h})}, \epsilon_3^{(g)}$ are all positive arbitrarily small constants, and $C_{\mathbf{w}_{up+}}$ is a constant such that $\|\mathbf{w}^+\|_2 \leq C_{\mathbf{w}_{up+}}$.

Proof. Follows from the previous discussion. $\square$

3.3 Matching upper and lower bounds

In this section we specialize the general bounds introduced above and show how they match. We will again divide presentation in three subsections. In the first of the subsections we will make a connection to the noiseless "signed" case and show how one can then remove the constraint from (184), (185), and (186). In the second subsection we will consider a $\mathbf{w}$ such that $|\|\mathbf{w}\|_2 - \|\widehat{\mathbf{w}^+}\|_2| \geq \epsilon_{\mathbf{w}_{up}}\|\widehat{\mathbf{w}^+}\|_2$. We will then quantify how much the lower bound that can be computed for such a $\mathbf{w}$ through the framework presented in Section 3.1 deviates from the optimal one obtained for $\widehat{\mathbf{w}^+}$. In the last subsection we will then show that there will be a $\mathbf{w}$ such that the upper bound computed through the framework presented in Section 3.2 will deviate less. That will in essence establish that upper and lower bounds computed in the previous sections indeed match.

3.3.1 Connection to the ℓ_1 optimization of signed $\mathbf{x}$

In this subsection we establish a connection between the constraint in (184), (185), and (186) and the fundamental performance characterization of ℓ_1 optimization derived in [62] (and of course earlier in the context of neighborly polytopes in [27]). We first recall on the condition from Lemma 11. The condition states

$$\sqrt{1 + \frac{\sigma^2}{C_{\mathbf{w}}^2}}\|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2 \leq \|\mathbf{g}\|_2, \tag{203}$$

where $C_{\mathbf{w}}$ is an arbitrarily large constant and $\widehat{\nu^+}$ and $\widehat{\lambda^{(2+)}}$ are the solutions of

$$\max \quad \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)}\tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n$$
$$\nu \geq 0. \tag{204}$$

Now we note the following equivalent to (204) for the case when nonzero components of $\tilde{\mathbf{x}}$ are infinite

$$\max \quad \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu\mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2}$$

$$\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n-k$$
$$\lambda_i^{(2)} = 0, n-k+1 \leq i \leq n$$
$$\nu \geq 0. \tag{205}$$

To make the new observations easily comparable to the corresponding ones from [61, 63] we set

$$\bar{\mathbf{h}}^+ = [\mathbf{h}_{(1)}^{(1)}, \mathbf{h}_{(2)}^{(2)}, \ldots, \mathbf{h}_{(n-k)}^{(n-k)}, \mathbf{h}_{n-k+1}, \mathbf{h}_{n-k+2}, \ldots, \mathbf{h}_n]^T, \tag{206}$$

where $[\mathbf{h}_{(1)}^{(1)}, \mathbf{h}_{(2)}^{(2)}, \ldots, \mathbf{h}_{(n-k)^{(n-k)}}]$ are $[\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_{n-k}]$ sorted in increasing order (possible ties in the sorting process are of course broken arbitrarily). Also we let $\mathbf{z}^{(2)}$ be as in the previous section, i.e. let it be such that $\mathbf{z}_i^{(2)} = -\mathbf{z}_i^{(1)}, n-k+1 \leq i \leq n$ and $\mathbf{z}_i^{(2)} = \mathbf{z}_i^{(1)}, 1 \leq i \leq n-k$. It is then relatively easy to see

that the above optimization problem is equivalent to

$$\max \quad \sigma\sqrt{\|\mathbf{g}\|_2^2 - \|\bar{\mathbf{h}}^+ - \nu\mathbf{z}^{(2)} + \lambda^{(2)}\|_2^2}$$
$$\text{subject to} \quad \lambda_i^{(2)} \geq 0, 1 \leq i \leq n - k$$
$$\lambda_i^{(2)} = 0, n - k + 1 \leq i \leq n$$
$$\nu \geq 0. \tag{207}$$

Let ν_{ℓ_1+} and $\lambda^{(\ell_1+)}$ be the solution of the above maximization. Further, consider the following "signed" version of the standard ℓ_1-optimization

$$\min \quad \|\mathbf{x}\|_1$$
$$\text{subject to} \quad A\mathbf{x} = \mathbf{y}$$
$$\mathbf{x}_i \geq 0. \tag{208}$$

Then, as we showed in [63] and [62], the inequality

$$E\|\mathbf{g}\|_2 > E\|\bar{\mathbf{h}}^+ - \nu_{\ell_1+}\mathbf{z}^{(2)} + \lambda^{(\ell_1+)}\|_2 \tag{209}$$

establishes the following fundamental performance characterization of the ℓ_1 optimization algorithm from (208) that could be used instead of LASSO from (158) to recover signed $\mathbf{x}$ in (1) (which is a noiseless version of (3))

$$(1 - \beta_w^+)\frac{\sqrt{\frac{1}{2\pi}}e^{-(\text{erfinv}(2\frac{1-\alpha_w^+}{1-\beta_w^+}-1))^2}}{\alpha_w^+} - \sqrt{2}\text{erfinv}(2\frac{1-\alpha_w^+}{1-\beta_w^+} - 1) = 0, \tag{210}$$

where of course $\alpha_w^+ = \frac{m}{n}$ and $\beta_w^+ = \frac{k}{n}$. As it is also shown in [63] and [62] both of the quantities under the expected values in (209) nicely concentrate. Then with overwhelming probability one has that for any pair (α, β) that satisfies (or lies below) the fundamental performance characterization of ℓ_1 optimization given in (210)

$$\|\mathbf{g}\|_2 > \|\bar{\mathbf{h}}^+ - \nu_{\ell_1+}\mathbf{z}^{(2)} + \lambda^{(\ell_1+)}\|_2. \tag{211}$$

Moreover, since $\lambda_i^{(2)} \geq 0, n - k + 1 \leq i \leq n$, in (204) one actually has that (211) implies that with overwhelming probability

$$\|\mathbf{g}\|_2 > \|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2, \tag{212}$$

which for sufficiently large $C_{\mathbf{w}}$ is the same as (203). We then in what follows assume that pair (α, β) is such that it satisfies the "signed" fundamental ℓ_1 optimization performance characterization (or is in the region below it) from (210) and therefore proceed by ignoring condition (203).

3.3.2 Deviation from the lower-bound

In this subsection we establish that $\|\mathbf{w}_{lasso+}\|_2$ (of course, $\mathbf{w}_{lasso+} = \widehat{\mathbf{x}^+} - \tilde{\mathbf{x}}$, where $\widehat{\mathbf{x}^+}$ is the solution of (158)) can not deviate substantially from $\|\widehat{\mathbf{w}^+}\|_2$ without substantially affecting the value of the lower bound on the objective in (158) that is derived in Section 3.1. To that end let us assume that there is a $\mathbf{w}_{off+}$ that is the solution of the LASSO from (158) (or to be slightly more precise that is such that $\widehat{\mathbf{x}^+} = \tilde{\mathbf{x}} + \mathbf{w}_{off+}$, where obviously $\widehat{\mathbf{x}^+}$ is the solution of (158)). Further, let $|\|\mathbf{w}_{off+}\|_2 - \|\widehat{\mathbf{w}^+}\|_2| \geq \epsilon_{\mathbf{w}_{up}}\|\widehat{\mathbf{w}^+}\|_2$, where $\epsilon_{\mathbf{w}_{up}}$ is an arbitrarily small constant.

One can then write a "signed" analogue to (112)

$$\xi_{off+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},\mathbf{w}_{off+}) = \max_{\lambda^{(2)}\in\Lambda^{(2+)},\nu\geq0}(\sqrt{\mathbf{w}_{off}^2+\sigma^2}\|\mathbf{g}\|_2 - \mathbf{w}_{off}\|\mathbf{h}+\nu\mathbf{z}^{(1)}-\lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i).$$

(213)

After repeating all the arguments between (112) and Lemma 9 one obtains the following analogue to Lemma 9.

Lemma 15. *Let $\mathbf{v}$ be an $n\times1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m\times n$ matrix of i.i.d. standard normal random variables. Consider an $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x}=\tilde{\mathbf{x}}$. Let then ζ_{obj+} be as defined in (159) and let $\mathbf{w}_{off+}$ be the solution of (162). Let α and β be below the fundamental characterization (210) and let $\widehat{\mathbf{w}^+}$ be as defined in (185). Assume that $|\|\mathbf{w}_{off+}\|_2 - \|\widehat{\mathbf{w}^+}\|_2| \geq \epsilon_{\mathbf{w}_{up}}\|\widehat{\mathbf{w}^+}\|_2$, where $\epsilon_{\mathbf{w}_{up}}$ is an arbitrarily small but fixed constant. Then there would be a constant $\epsilon_{off}>0$, and arbitrarily small positive constants $\epsilon_{lip},\epsilon_1^{(\mathbf{h})},\epsilon_1^{(g)}$ such that*

$$P(\zeta_{obj+}\geq\zeta_{obj+}^{(off)})\geq1-e^{-\epsilon_{off}n},$$

(214)

where

$$\zeta_{obj}^{(off)} = (1-\epsilon_{lip})(1+\frac{\epsilon_{\mathbf{w}_{up}}^2}{2(1+\epsilon_{\mathbf{w}_{up}})})E\xi_{ov+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}) - \epsilon_1^{(\mathbf{h})}\sqrt{n} - \epsilon_1^{(g)}\sqrt{n},$$

(215)

and $\xi_{ov+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}})$ is as defined in (181).

Proof. Follows from the discussion in Section 2.3.2. □

3.3.3 Deviation of the upper bound

In this section we establish that $\|\mathbf{w}_{lasso+}\|_2$ can not deviate from $\|\widehat{\mathbf{w}^+}\|_2$ as much as it was assumed in the previous section which is conceptually enough to make the bounds from Sections 3.1 and 3.2 match. All arguments from Section 2.3.3 can be repeated again. The only difference will be that in all optimization problems from Section 2.3.3 one will now have no upper bound on $\lambda_i^{(2)},1\leq i\leq n$ (this essentially amounts to using set $\Lambda^{(2+)}$ instead of set $\Lambda^{(2)}$). One then has a "signed" analogue to (142)

$$E\xi_{up+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}},E\|\widehat{\mathbf{w}^+}\|_2) = E\max_{\lambda^{(2)}\in\Lambda^{(2+)},\nu\geq0}(\sqrt{(E\|\widehat{\mathbf{w}^+}\|_2)^2+\sigma^2}\|\mathbf{g}\|_2 - E\|\widehat{\mathbf{w}^+}\|_2\|\mathbf{h}+\nu\mathbf{z}^{(1)}-\lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i)$$

$$= E\max_{\lambda^{(2)}\in\Lambda^{(2+)},\nu\geq0}(\sqrt{(E\widehat{d^+})^2+\sigma^2}\|\mathbf{g}\|_2 - E\widehat{d^+}\|\mathbf{h}+\nu\mathbf{z}^{(1)}-\lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i)$$

$$\doteq E\min_{d\geq0}\max_{\lambda^{(2)}\in\Lambda^{(2+)},\nu\geq0}(\sqrt{d^2+\sigma^2}\|\mathbf{g}\|_2 - d\|\mathbf{h}+\nu\mathbf{z}^{(1)}-\lambda^{(2)})\|_2 - \sum_{i=n-k+1}^{n}\lambda_i^{(2)}\tilde{\mathbf{x}}_i) = E\xi_{ov+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}}),$$

(216)

where $\widehat{d^+}=\|\widehat{\mathbf{w}^+}\|_2$ would be the solution of a "signed" analogue to (125). Following the arguments after (142) one then has that the assumption of Lemma 13 is unsustainable and that $\|\mathbf{w}_{lasso+}\|_2$ can not deviate substantially from $\|\widehat{\mathbf{w}^+}\|_2$. This then implies that with overwhelming probability the objective value of (158) concentrates around $E\xi_{ov+}(\sigma,\mathbf{g},\mathbf{h},\tilde{\mathbf{x}})$ and consequently that $\|\mathbf{w}_{lasso+}\|_2$ concentrates around $E\|\widehat{\mathbf{w}^+}\|_2$.

3.4 Connecting all pieces

In this section we connect all of the above. The following theorem essentially does so.

Theorem 3. *Let $\mathbf{v}$ be an $n \times 1$ vector of i.i.d. zero-mean variance σ^2 Gaussian random variables and let A be an $m \times n$ matrix of i.i.d. standard normal random variables. Further, let $\mathbf{g}$ and $\mathbf{h}$ be $m \times 1$ and $n \times 1$ vectors of i.i.d. standard normals, respectively. Consider a k-sparse $\tilde{\mathbf{x}}$ defined in (7) and a $\mathbf{y}$ defined in (3) for $\mathbf{x} = \tilde{\mathbf{x}}$. Let the solution of (158) be $\widehat{\mathbf{x}^+}$ and let the so-called error vector of LASSO from (158) be $\mathbf{w}_{lasso+} = \widehat{\mathbf{x}^+} - \tilde{\mathbf{x}}$. Let n be large and let constants $\alpha = \frac{m}{n}$ and $\beta = \frac{k}{n}$ be below the fundamental characterization (210). Consider the following optimization problem:*

$$\xi_{ov+}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \max_{\nu, \lambda^{(2)}} \quad \sigma \sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2^2} - \sum_{i=n-k+1}^{n} \lambda_i^{(2)} \tilde{\mathbf{x}}_i$$

$$\text{subject to} \quad \nu \geq 0$$

$$\lambda_i^{(2)} \geq 0, 1 \leq i \leq n. \tag{217}$$

Let $\widehat{\nu^+}$ and $\widehat{\lambda^{(2+)}}$ be the solution of (217). Set

$$\|\widehat{\mathbf{w}^+}\|_2 = \sigma \frac{\|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \widehat{\nu^+}\mathbf{z}^{(1)} - \widehat{\lambda^{(2+)}}\|_2^2}}. \tag{218}$$

Then:

$$P((1-\epsilon_1^{(lasso)})E\xi_{ov+}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \leq \|\mathbf{y} - A\hat{\mathbf{x}}\|_2 \leq (1+\epsilon_1^{(lasso)})E\xi_{ov+}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}})) = 1 - e^{-\epsilon_2^{(lasso)}n} \tag{219}$$

and

$$P((1 - \epsilon_1^{(lasso)})E\|\widehat{\mathbf{w}^+}\|_2 \leq \|\mathbf{w}_{lasso+}\|_2 \leq (1 + \epsilon_1^{(lasso)})E\|\widehat{\mathbf{w}^+}\|_2) = 1 - e^{-\epsilon_2^{(lasso)}n}, \tag{220}$$

where $\epsilon_1^{(lasso)} > 0$ is an arbitrarily small constant and $\epsilon_2^{(lasso)}$ is a constant dependent on $\epsilon_1^{(lasso)}$ and σ but independent of n.

Proof. Follows from the above discussion. $\qquad \square$

3.4.1 LASSO's generic performance

In this section we show how the results presented in the above theorem can be adapted to the so-called "worst-case" scenario or as we refer to it "generic performance" scenario. Repeating the line of arguments from Section 2.4.1 one can establish the following *generic* equivalent to Theorem 3.

Theorem 4. *Assume the setup of Theorem 3. Consider the following optimization problem:*

$$\xi_{ov+}^{(gob)}(\sigma, \mathbf{g}, \mathbf{h}) = \min_{\nu, \lambda^{(2)}} \quad \|\mathbf{h} + \nu \mathbf{z}^{(1)} - \lambda^{(2)}\|_2$$

$$\text{subject to} \quad \nu \geq 0$$

$$\lambda_i^{(2)} = 0, n - k + 1 \leq i \leq n$$

$$\lambda_i^{(2)} \geq 0, 1 \leq i \leq n - k. \tag{221}$$

Let ν_{gen+} and $\lambda^{(gen+)}$ be the solution of (221). Set

$$\|\mathbf{w}_{gen+}\|_2 = \sigma \frac{\|\mathbf{h} + \nu_{gen+}\mathbf{z}^{(1)} - \lambda^{(gen+)}\|_2}{\sqrt{\|\mathbf{g}\|_2^2 - \|\mathbf{h} + \nu_{gen+}\mathbf{z}^{(1)} - \lambda^{(gen+)}\|_2^2}}. \tag{222}$$

Then:

$$\begin{aligned} P(\exists\mathbf{w}_{lasso+}|\|\mathbf{w}_{lasso+}\|_2 \in ((1 - \epsilon_1^{(lasso)})E\|\mathbf{w}_{gen+}\|_2, (1 + \epsilon_1^{(lasso)})E\|\mathbf{w}_{gen+}\|_2)) &\geq 1 - e^{-\epsilon_2^{(lasso)}n} \\ P(\|\mathbf{w}_{lasso+}\|_2 \leq (1 + \epsilon_1^{(lasso)})E\|\mathbf{w}_{gen+}\|_2) &\geq 1 - e^{-\epsilon_3^{(lasso)}n}, \end{aligned} \tag{223}$$

where $\epsilon_1^{(lasso)} > 0$ is an arbitrarily small constant and $\epsilon_2^{(lasso)}$ and $\epsilon_3^{(lasso)}$ are constants dependent on $\epsilon_1^{(lasso)}$ and σ but independent of n.

Proof. Follows by the use of the same arguments that were used to establish Theorem 3. $\qquad\square$

The following corollary then provides a quick way of computing the concentrating point of the "worst case" norm of the error vector.

Corollary 2. *Assume the setup of Theorems 3 and 4. Let $\alpha = \frac{m}{n}$ and $\beta_w^+ = \frac{k}{n}$. Then*

$$\begin{aligned} P(\exists\mathbf{w}_{lasso+}|\|\mathbf{w}_{lasso+}\|_2 \in ((1 - \epsilon_1^{(lasso)})\sigma\sqrt{\frac{\alpha_w^+}{\alpha - \alpha_w^+}}, (1 + \epsilon_1^{(lasso)})\sigma\sqrt{\frac{\alpha_w^+}{\alpha - \alpha_w^+}})) &\geq 1 - e^{-\epsilon_2^{(lasso)}n} \\ P(\|\mathbf{w}_{lasso+}\|_2 \leq (1 + \epsilon_1^{(lasso)})\sigma\sqrt{\frac{\alpha_w^+}{\alpha - \alpha_w^+}}) &\geq 1 - e^{-\epsilon_3^{(lasso)}n}, \end{aligned} \tag{224}$$

where $\alpha_w^+ < \alpha$ is such that

$$(1 - \beta_w^+)\frac{\sqrt{\frac{1}{2\pi}}e^{-(\mathit{erfinv}(2\frac{1-\alpha_w^+}{1-\beta_w^+}-1))^2}}{\alpha_w^+} - \sqrt{2}\mathit{erfinv}(2\frac{1-\alpha_w^+}{1-\beta_w^+} - 1) = 0, \tag{225}$$

$\epsilon_1^{(lasso)} > 0$ is an arbitrarily small constant, and $\epsilon_2^{(lasso)}$ and $\epsilon_2^{(lasso)}$ are constants dependent on $\epsilon_1^{(lasso)}$ and σ but independent of n.

Proof. Follows by the use of the same arguments that were used to establish Corollary 1 and a recognition that the fundamental characterization of interest in the "signed" case is the one given in (210). $\qquad\square$

Based on the above corollary one can then for any σ and any pair (α, β_w^+) (that is below fundamental characterization (225) or alternatively (210)) determine the value of the worst case $E\|\mathbf{w}_{lasso+}\|_2$ as $\sigma\sqrt{\frac{\alpha_w^+}{\alpha - \alpha_w^+}}$. We present the obtained results in Figure 4. For several fixed values of the worst case $E\|\mathbf{w}_{lasso+}\|_2$ we determine curves of points (α, β_w^+) for which these fixed values are achieved (of course for any α that is below a curve the value of the corresponding worst case $E\|\mathbf{w}_{lasso+}\|_2$ is smaller). As in the previous section, the lower the norm-2 of the error vector the smaller the allowable region for pairs (α, β_w^+). Also as it was the case in the previous section, the results of the above corollary match those obtained in [7, 26] through a state evolution/bilief propagation type of analysis for the "signed" version of

the LASSO from (5) (signed version of the LASSO from (5) as expected assumes just simple adding of the positivity constraints on the components of $\mathbf{x}$).
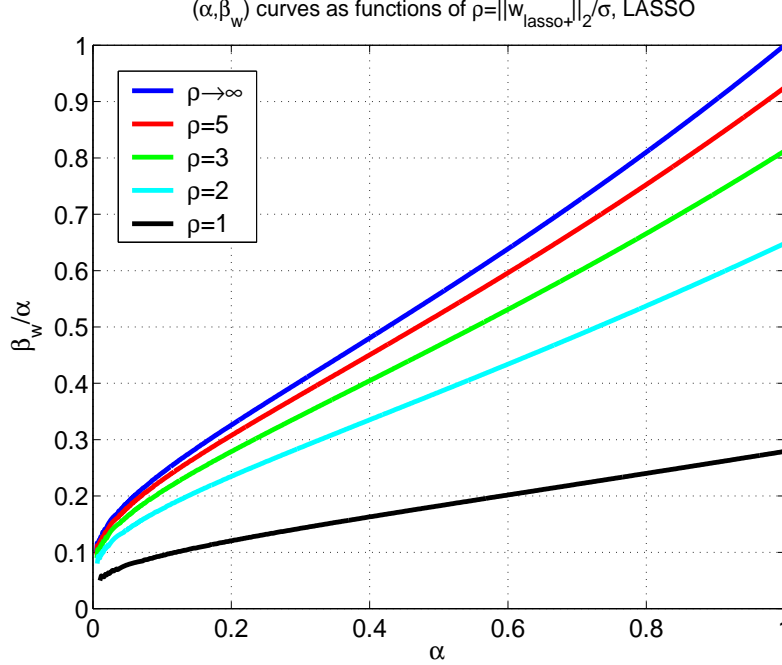

Figure 4: (α, β_w^+) curves as functions of $\rho = \frac{E\|\mathbf{w}_{lasso+}\|_2}{\sigma}$ for LASSO algorithm from (158)

4 Connecting LASSO's from (6) and (5)

In this section we establish a connection between the LASSO algorithm from (6) that we analyzed in Section 2 and the more well known form of LASSO from (5). Instead of well-known (5) we will consider its a slight modification

$$\min_{\mathbf{x}} \|\mathbf{y} - A\mathbf{x}\|_2 + \lambda_{lasso}\|\mathbf{x}\|_1. \tag{226}$$

Both LASSO's, (6) and (226), (as well as the one from (5)) rely on some type of the prior knowledge that can be available about A, $\mathbf{v}$, or $\tilde{\mathbf{x}}$. In (6) we assumed that one knows $\|\tilde{\mathbf{x}}\|_1$ (of course, if one has no knowledge of $\|\tilde{\mathbf{x}}\|_1$ LASSO from (6) simply can not be run). On the other hand the LASSO from (226) (as well as the one from (5)) requires that one sets in advance parameter λ_{lasso} which can be a tough task if there is no a priori knowledge about A, $\mathbf{v}$, or $\tilde{\mathbf{x}}$. Now even if there is some a priori available knowledge about these objects there are still many ways how one can set λ_{lasso}. We will below show a particular way of setting λ_{lasso} in (226) that can make LASSO's from (6) and (226) essentially equivalent (of course, as long as one is interested in performance measures discussed in this paper). In the interest of saving space we will sketch only the key arguments without going into tedious details similar to the ones presented in earlier sections. (All that we mention below can be made precise, though; in fact, one can pretty much reach the same level of exactness demonstrated in Sections 2 and 3; however, the length of the precise probabilistic arguments would equal (if not exceed) the length of the arguments presented in Sections 2 and 3.)

Now, let λ_{lasso} in (226) be such that $\lambda_{lasso} = E\hat{\nu}$ where $\hat{\nu}$ is the solution of (42). Then (226) becomes

$$\min_{\mathbf{x}} \|\mathbf{y} - A\mathbf{x}\|_2 + E\hat{\nu}\|\mathbf{x}\|_1, \tag{227}$$

or in a more convenient form

$$\min_{\mathbf{x}} \|\mathbf{y} - A\mathbf{x}\|_2 + E\hat{\nu}\|\mathbf{x}\|_1 - E\hat{\nu}\|\tilde{\mathbf{x}}\|_1, \tag{228}$$

This could be rewritten in a way analogous to (14) as

$$\min_{\mathbf{w}} \quad \|A_{\mathbf{v}} \begin{bmatrix} \mathbf{w} \\ \sigma \end{bmatrix} \|_2 + E\hat{\nu}\|\tilde{\mathbf{x}} + \mathbf{w}\|_1 - E\hat{\nu}\|\tilde{\mathbf{x}}\|_1, \tag{229}$$

where $A_{\mathbf{v}}$ is as in (14). One can then repeat all arguments from the beginning of Section 2.1 (essentially those before Section 2.1.1) to arrive at the following analogue of (23)

$$\xi_{conn}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) = \min_{\mathbf{w}} \quad \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2}\|\mathbf{g}\|_2 + \sum_{i=1}^{n} \mathbf{h}_i \mathbf{w}_i + E\hat{\nu}\|\tilde{\mathbf{x}} + \mathbf{w}\|_1 - E\hat{\nu}\|\tilde{\mathbf{x}}\|_1$$

$$\text{subject to} \quad \sqrt{\|\mathbf{w}\|_2^2 + \sigma^2} \leq \sqrt{C_{\mathbf{w}}^2 + \sigma^2}. \tag{230}$$

Now, one should note that $E\hat{\nu}$ in the above optimization is chosen as the "optimal" (it is actually the concentrating point of the optimal one; to make this really precise one would need to go through all the probabilistic arguments of Section 2 and plus some more) ν in the Lagrange dual of (23). One then has that arguments from Section 2.1 (essentially an appropriate repetition of those that follow (23)) will produce the lower bound on the objective of (228) that is with overwhelming probability arbitrarily close to the one derived in Lemma 5. The arguments from Section 2.2 related to the upper bound can be trivially repeated as well since the negativity of the objective in (67) implies that r is also an upper bound on the optimal value of the objective in (228). The matching arguments from Section 2.3 then follow as well. Now if one let $\mathbf{w}_{conn}$ be the solution of (229), then with overwhelming probability $\|\mathbf{w}_{conn}\|_2$ concentrates around $E\|\hat{\mathbf{w}}\|_2$, where $\hat{\mathbf{w}}$ is as defined in Theorem 1.

For the signed case the arguments are the same, only instead of $E\hat{\nu}$ in (227), (228), and (229) one should use $E\widehat{\nu^+}$ where $\widehat{\nu^+}$ is the solution of (181). Also, as it is probably obvious, this time $\|\mathbf{w}_{conn}\|_2$ concentrates around $E\|\widehat{\mathbf{w}^+}\|_2$ where $\widehat{\mathbf{w}^+}$ is as defined in Theorem 3.

5 A relation between a LASSO and an SOCP

In this section we show that there is an SOCP equivalent to the LASSO from (6) (as long as the norm-2 of the error vector is a performance measure of interest). To that end let us recall that an SOCP algorithm for finding an approximation of $\tilde{\mathbf{x}}$ if A and $\mathbf{y}$ from (3) are known can be (see, e.g. [13])

$$\min_{\mathbf{x}} \quad \|\mathbf{x}\|_1$$

$$\text{subject to} \quad \|\mathbf{y} - A\mathbf{x}\|_2 \leq r_{socp}. \tag{231}$$

The choice of r_{socp} critically impacts the outcome of the above optimization. In fact more is true, the choice of r_{socp} heavily depends on what type of approximation error one is looking for. As we have mentioned in Section 1 a popular choice for r_{socp} is the smallest quantity that is with high probability larger than $\|\mathbf{v}\|_2$. There are probably many reasons for such a choice. One of them is that it would with high probability guarantee that the original $\tilde{\mathbf{x}}$ in (3) is permissible in (231). Now, if one is looking for an $\mathbf{x}$ that will be close in norm-2 to the original $\tilde{\mathbf{x}}$ then it is not necessary to look for the original $\tilde{\mathbf{x}}$ (especially so given that finding original $\tilde{\mathbf{x}}$ is in general pretty much impossible). So if one gives up on that then the value of r_{socp} can go even lower than the smallest quantity larger (with overwhelming probability) than $\|\mathbf{v}\|_2$. One should also

note that by lowering r_{socp} one would give up not only possibility to find $\tilde{\mathbf{x}}$ (which is tiny anyway) but also highly likely more when it comes to the structure of the solution vector. This is of course a problem on its own that requires a thorough discussion. However, since we now look only at the norm of the error vector as a performance measure we stop short of pursuing this discussion here any further.

To go along these lines we choose

$$r_{socp} = E\xi_{up}(\sigma, \mathbf{g}, \mathbf{h}, E\|\hat{\mathbf{w}}\|_2) \doteq E\xi_{ov}(\sigma, \mathbf{g}, \mathbf{h}, \tilde{\mathbf{x}}) \leq E\xi_{ov}^{(gob)}(\sigma, \mathbf{g}, \mathbf{h}) \leq \sigma\sqrt{m} \doteq E\|\mathbf{v}\|_2. \qquad (232)$$

Now, let $\mathbf{x}_{socp}$ be the solution of (231) (with r_{socp} as in (232)). Further let $\mathbf{x}_{socp} = \mathbf{w}_{socp} + \tilde{\mathbf{x}}$. Let $\hat{\mathbf{w}}$ be as defined in Theorem 1. Then as shown in Sections 2.1, 2.2, and 2.3 $\|\mathbf{w}_{socp}\|_2$ concentrates around $E\|\hat{\mathbf{w}}\|_2$. Basically, the argument is that if $|\|\mathbf{w}_{socp}\|_2 - \|\hat{\mathbf{w}}\|_2| \geq \epsilon_{\mathbf{w}_{up}}\|\hat{\mathbf{w}}\|_2$ for a fixed arbitrarily small positive $\epsilon_{\mathbf{w}_{up}}$ and $\|\tilde{\mathbf{x}} + \mathbf{w}_{socp}\|_1 \leq \|\tilde{\mathbf{x}}\|_1$ then with overwhelming probability $\|\mathbf{y} - A\mathbf{x}_{socp}\|_2 > r_{socp}$. On the other hand, as it was also established in Sections 2.1, 2.2, and 2.3, there is a $\mathbf{w}$ (which concentrates around $E\|\hat{\mathbf{w}}\|_2$ with overwhelming probability) such that $\|\tilde{\mathbf{x}} + \mathbf{w}\|_1 \leq \|\tilde{\mathbf{x}}\|_1$ and $\|\mathbf{y} - A\mathbf{x}_{socp}\|_2 \leq r_{socp}$. This essentially establishes that if one chooses r_{socp} in (231) as suggested in (232) then the norm-2 of the error vector will be the same as the norm-2 of the error vector one obtains through LASSO's from (6) and (226) (the latter one of course with an appropriate choice of λ_{lasso}).

As we hinted above what we presented here is only a characterization of a particular performance measure of an SOCP algorithm (the same is of course true for the LASSO algorithms). How adequate is such a performance measure is whole another story that we will explore in more detail elsewhere.

6   Numerical results

In this section we present a set of numerical results related to the theoretical predictions that we derived in earlier sections. We will divide the presentation into two groups: 1) the set of results that will relate to the general (unsigned) unknown sparse vectors and 2) the set of results that will relate to signed unknown sparse vectors. To make scaling easier in all experiments we set $\sigma = 1$. We also assumed that nonzero components of $\tilde{\mathbf{x}}$ are all of equal and large magnitude. For the concreteness we set this magnitude to be $\frac{1000}{\sqrt{n}}$. For every setup that we discuss below we ran 100 numerical experiments.

6.1   Numerical results related to general $\mathbf{x}$

In this subsection we will present numerical results that relate to the theoretical ones created in Sections 2 and 4. We will consider two groups of (α, β_w) regimes, one that we will refer to as the low (α, β_w) regime and the other that we will refer to as the high (α, β_w) regime.

2) Low (α, β_w) *regime* — $\rho = \frac{E\|\mathbf{w}_{lasso}\|_2}{\sigma} = 2$

We ran a carefully designed set of experiments intended to show a specific behavior of the LASSO's from (6) and (227) in what we will refer to as the low (α, β_w) regime. For $\alpha \in \{0.3, 0.5, 0.7\}$ we determined three values of β_w from the contour LASSO line that corresponds to $\rho = 2$ in the figure given in Section 2. We then ran (6) assuming that $\|\tilde{\mathbf{x}}\|_1$ is known and (227) using theoretical value for $E\hat{\nu}$ where, as mentioned in Section 4, $\hat{\nu}$ is the solution of (42). We call the optimal value of the objective in (228) ζ_{conn} (this value is the optimal value of (227) shifted by a constant). Also, for this set of experiments we set $n = 2000$. Obtained results are presented in Table 1. The theoretical values for any of the simulated quantities in any of the simulated scenarios are given in parallel as bolded numbers. We observe a solid agreement between the theoretical predictions and the results obtained through numerical experiments.

2) High (α, β_w) *regime* — $\rho = \frac{E\|\mathbf{w}_{lasso}\|_2}{\sigma} = 3$

Table 1: Experimental/**theoretical** results for the noisy recovery through LASSO's; $\sigma = 1$, $\rho = E\|\mathbf{w}_{lasso}\|_2 = 2$; (6) and (227) were run 100 times with $n = 2000$

α	β_w/α	$E\hat{\nu}$	$\frac{E\zeta_{conn}}{\sqrt{n}}$	$E\|\mathbf{w}_{conn}\|_2$	$\frac{E\zeta_{obj}}{\sqrt{n}}$	$E\|\mathbf{w}_{lasso}\|_2$
0.3	0.21	**1.3141**	0.2444/**0.2449**	2.0225/**2**	0.2449/**0.2449**	2.0188/**2**
0.5	0.27	**1.0227**	0.3159/**0.3162**	2.0058/**2**	0.3162/**0.3162**	2.0018/**2**
0.7	0.33	**0.7959**	0.3717/**0.3742**	2.0168/**2**	0.3721/**0.3742**	2.0155/**2**

Table 2: Experimental/**theoretical** results for the noisy recovery through LASSO's; $\sigma = 1$, $\rho = E\|\mathbf{w}_{lasso}\|_2 = 3$; (6) and (227) were run 100 times

α	β_w/α	$E\hat{\nu}$	$\frac{E\zeta_{conn}}{\sqrt{n}}$	$E\|\mathbf{w}_{conn}\|_2$	$\frac{E\zeta_{obj}}{\sqrt{n}}$	$E\|\mathbf{w}_{lasso}\|_2$
0.3	0.249	**1.2508**	0.1699/**0.1732**	3.1714/**3**	0.1705/**0.1732**	3.1507/**3**
0.5	0.325	**0.9477**	0.2231/**0.2236**	3.0560/**3**	0.2239/**0.2236**	3.0405/**3**
0.7	0.41	**0.7046**	0.2579/**0.2646**	3.1166/**3**	0.2585/**0.2646**	3.1069/**3**

We also ran a carefully designed set of experiments intended to show a specific behavior of the LASSO's from (6) and (227) in what we will refer to as the high (α, β_w) regime. For $\alpha \in \{0.3, 0.5, 0.7\}$ we now determined three values of β_w from the contour LASSO line that corresponds to $\rho = 3$ in the figure given in Section 2. We then again ran (6) assuming that $\|\tilde{\mathbf{x}}\|_1$ is known and (227) using the theoretical values for $E\hat{\nu}$. For the scenario when $\alpha = 0.3$ we set $n = 3000$ while for the scenarios with other two values of α we set $n = 2000$. Obtained results are presented in Table 2. The theoretical values for any of the simulated quantities in any of the simulated scenarios are again given in parallel as bolded numbers. We again observe a solid agreement between the theoretical predictions and the results obtained through numerical experiments.

6.2 Numerical results related to signed x

In this subsection we will present numerical results that relate to the theoretical ones created in Sections 3 and 4. We will again consider two groups of (α, β_w^+) regimes, one that we will refer to as the low (α, β_w^+) regime and the other that we will refer to as the high (α, β_w^+) regime.

  2) Low (α, β_w^+) regime — $\rho = \frac{E\|\mathbf{w}_{lasso+}\|_2}{\sigma} = 2$

We first ran a set of experiments intended to show a specific behavior of the LASSO's from (158) and (227) in what we will refer to as the low (α, β_w^+) regime. For $\alpha \in \{0.3, 0.5, 0.7\}$ we determined three values of β_w^+ from the contour LASSO line that corresponds to $\rho = 2$ in the figure given in Section 3. We then ran (158) assuming that $\|\tilde{\mathbf{x}}\|_1$ is known and (227) using theoretical value for $\widehat{E\nu^+}$ where, as mentioned in Section 4, $\widehat{\nu^+}$ is the solution of (181). Also when running (227) we now added positivity constraints on the elements of $\mathbf{x}$. We call the optimal value of the objective in (228) ζ_{conn+}. When $\alpha = 0.7$ we set $n = 1500$ while for the other two values of α we set $n = 2000$. Obtained results are presented in Table 3. The theoretical values for any of the simulated quantities in any of the simulated scenarios are as usual given in parallel as bolded numbers. We once again observe a solid agreement between the theoretical predictions and the results obtained through numerical experiments.

  2) High (α, β_w^+) regime — $\rho = \frac{E\|\mathbf{w}_{lasso+}\|_2}{\sigma} = 3$

As in the previous subsection we also ran a carefully designed set of experiments intended to show a specific behavior of the LASSO's from (158) and (227) in what we will refer to as the high (α, β_w^+) regime.

Table 3: Experimental/**theoretical** results for the noisy recovery through LASSO's; $\sigma = 1$, $\rho = E\|\mathbf{w}_{lasso+}\|_2 = 2$; (6) and (227) were run 100 times

α	β_w^+/α	$E\widehat{\nu^+}$	$\frac{E\zeta_{conn+}}{\sqrt{n}}$	$E\|\mathbf{w}_{conn+}\|_2$	$\frac{E\zeta_{obj+}}{\sqrt{n}}$	$E\|\mathbf{w}_{lasso+}\|_2$
0.3	0.286	**0.9592**	0.2454/**0.2449**	1.9939/**2**	0.2461/**0.2449**	1.9876/**2**
0.5	0.3842	**0.6516**	0.3133/**0.3162**	2.0229/**2**	0.3140/**0.3162**	2.0177/**2**
0.7	0.4849	**0.4292**	0.3786/**0.3742**	1.9947/**2**	0.3794/**0.3742**	1.9886/**2**

Table 4: Experimental/**theoretical** results for the noisy recovery through LASSO's; $\sigma = 1$, $\rho = E\|\mathbf{w}_{lasso+}\|_2 = 3$; (6) and (227) were run 100 times

α	β_w^+/α	$E\widehat{\nu^+}$	$\frac{E\zeta_{conn+}}{\sqrt{n}}$	$E\|\mathbf{w}_{conn+}\|_2$	$\frac{E\zeta_{obj+}}{\sqrt{n}}$	$E\|\mathbf{w}_{lasso+}\|_2$
0.3	0.3423	**0.8197**	0.1713/**0.1732**	3.1213/**3**	0.1723/**0.1732**	3.0898/**3**
0.5	0.4672	**0.5757**	0.2245/**0.2236**	2.9983/**3**	0.2255/**0.2236**	2.9860/**3**
0.7	0.5971	**0.3470**	0.2644/**0.2646**	3.0373/**3**	0.2654/**0.2646**	3.0218/**3**

Following further the methodology of the previous subsection for $\alpha \in \{0.3, 0.5, 0.7\}$ we determined three values of β_w^+ from the contour LASSO line that corresponds to $\rho = 3$ in the figure given in Section 3. We then again ran (158) and (227) (when running (227) we of course again added positivity constraints and we again used theoretical value for $E\widehat{\nu^+}$). When $\alpha = 0.7$ we set $n = 1500$ while for the other two values of α we set $n = 2000$. Obtained results are presented in Table 4. The theoretical values for all quantities of interest are again given in parallel as bolded numbers. We once again observe a solid agreement between the theoretical predictions and the results obtained through numerical experiments.

7 Discussion

In this paper we considered "noisy" under-determined systems of linear equations with sparse solutions. We looked from a theoretical point of view at classical polynomial-time LASSO algorithms. Under the assumption that the system matrix A has i.i.d. standard normal components, we created a general framework that can be used to characterize various quantities of interest in analyzing the LASSO's performance. Among other things, the framework enables one to precisely estimate the norm of the error vector in "noisy" under-determined systems. Moreover, it can do so for any given k-sparse vector $\tilde{\mathbf{x}}$.

While many quantities of interest in LASSO recovery can be computed through the mechanism presented here, to demonstrate its power we in this introductory paper focused only on, what we called, LASSO's *generic* performance. We essentially established the precise values of the "worst-case" norm-2 of the error vector. On the other hand, using the framework one can create a massive set of results for the LASSO's non-generic or as we will refer to it *problem dependent* performance. However, this goes significantly over the scope of an introductory paper. We will dissect problems from this direction into tiny details in one of the forthcoming papers. Also, the existence of an SOCP type of the recovery algorithm that achieves the same norm-2 of the error vector as the LASSO does followed as a by-product of our analysis.

As for the applications, further developments are pretty much unlimited. Literally every problem that we were able to solve in the so-called noiseless case (and there was hardly any that we were not) through the mechanisms from [63] and [62] can now be handled in the noisy case as well. For example, quantifying performance of LASSO or SOCP optimization problems in solving "noisy" systems with special structure of the solution vector (block-sparse, binary, box-constrained, low-rank matrix, partially known locations of

nonzero components, just to name a few), "noisy" systems with noisy (or approximately sparse)) solution vectors can then easily be handled to an ultimate precision. In a series of forthcoming papers we will present some of these applications.

References

[1] R. Adamczak, A. E. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling. *Preprint*, 2009. available at arXiv:0904.4723.

[2] M. Akcakaya and V. Tarokh. A frame construction and a universal distortion bound for sparse representations. *IEEE Trans. on Signal Processing*, 56(6), June 2008.

[3] M. S. Asif and J. Romberg. On the lasso and dantzig selector equivalence. *44th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6, March 2010.

[4] R. Baraniuk, V. Cevher, M. Duarte, and C. Hegde. Model-based compressive sensing. available online at http://www.dsp.ece.rice.edu/cs/.

[5] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3), 2008.

[6] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *Preprint*. available online at arXiv:1001.3448.

[7] M. Bayati and A. Montanari. The lasso risk of gaussian matrices. *Preprint*. available online at arXiv:1008.2581.

[8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[9] F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Sparsity oracle inequalities for the lasso. *Electronic Journal of Statistics*, 1:169–194, 2007.

[10] E. Candes. Compressive sampling. *Proc. International Congress of Mathematics*, pages 1433–1452, 2006.

[11] E. Candes. The restricted isometry property and its implications for compressed sensing. *Compte Rendus de l'Academie des Sciences, Paris, Series I, 346*, pages 589–59, 2008.

[12] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. on Information Theory*, 52:489–509, December 2006.

[13] E. Candes, J. Romberg, and T. Tao. Stable signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59:1207–1223, 2006.

[14] E. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. on Information Theory*, 51:4203–4215, Dec. 2005.

[15] E. Candes, M. Wakin, and S. Boyd. Enhancing sparsity by reweighted l1 minimization. *J. Fourier Anal. Appl.*, 14:877–905, 2008.

[16] E. Cands and T. Tao. The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2313–2351, 2007.

[17] S.S. Chen and D. Donoho. Examples of basis pursuit. *Proceeding of wavelet applications in signal and image processing III*, 1995.

[18] S.S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM, JOurnal on Scientific Computing*, 20:33–61, 1998.

[19] S. Chretien. An alternating ell-1 approach to the compressed sensing problem. 2008. available online at http://www.dsp.ece.rice.edu/cs/.

[20] B. S. Cirelson, I. A. Ibragimov, and V. N. Sudakov. Norms of gaussian sample functions. *Lect. Notes Math.*, 50, 1976.

[21] G. Cormode and S. Muthukrishnan. Combinatorial algorithms for compressed sensing. *SIROCCO, 13th Colloquium on Structural Information and Communication Complexity*, pages 280–294, 2006.

[22] S. F. Cotter and B. D. Rao. Sparse channel estimation via matching pursuit with application to equalization. *IEEE Trans. on Communications*, 50(3), 2002.

[23] W. Dai and O. Milenkovic. Subspace pursuit for compressive sensing signal reconstruction. *Preprint*, page available at arXiv:0803.0811, March 2008.

[24] M. E. Davies and R. Gribonval. Restricted isometry constants where ell-p sparse recovery can fail for $0 < p \leq 1$. available online at http://www.dsp.ece.rice.edu/cs/.

[25] D. Donoho. High-dimensional centrally symmetric polytopes with neighborlines proportional to dimension. *Disc. Comput. Geometry*, 35(4):617–652, 2006.

[26] D. Donoho, A. Maleki, and A. Montanari. The noise-sensitiviy thase transition in compressed sensing. *Preprint*, Apr. 2010. available on arXiv.

[27] D. Donoho and J. Tanner. Neighborliness of randomly-projected simplices in high dimensions. *Proc. National Academy of Sciences*, 102(27):9452–9457, 2005.

[28] D. L. Donoho, M. Elad, and V. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, Jan 2006.

[29] D. L. Donoho, Y. Tsaig, I. Drori, and J.L. Starck. Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit. *2007*. available online at http://www.dsp.ece.rice.edu/cs/.

[30] M. Duarte, M. Davenport, D. Takhar, J. Laska, T. Sun, K. Kelly, and R. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2), 2008.

[31] B. Efron, T. Hastie, and R. Tibshirani. Discussion: The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2358–2364, 2007.

[32] S. Foucart and M. J. Lai. Sparsest solutions of underdetermined linear systems via ell-q minimization for $0 < q \leq 1$. available online at http://www.dsp.ece.rice.edu/cs/.

[33] M. P. Friedlander and M. A. Saunders. Discussion: The dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, 35(6):2385–2391, 2007.

[34] A. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin. Algorithmic linear dimension reduction in the l1 norm for sparse vectors. *44th Annual Allerton Conference on Communication, Control, and Computing*, 2006.

[35] A. Gilbert, M. J. Strauss, J. A. Tropp, and R. Vershynin. One sketch for all: fast algorithms for compressed sensing. *ACM STOC*, pages 237–246, 2007.

[36] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in R^n. *Geometric Aspect of of functional analysis, Isr. Semin. 1986-87, Lect. Notes Math*, 1317, 1988.

[37] R. Gribonval and M. Nielsen. Sparse representations in unions of bases. *IEEE Trans. Inform. Theory*, 49(12):3320–3325, December 2003.

[38] R. Gribonval and M. Nielsen. On the strong uniqueness of highly sparse expansions from redundant dictionaries. In *Proc. Int Conf. Independent Component Analysis (ICA'04)*, LNCS. Springer-Verlag, September 2004.

[39] R. Gribonval and M. Nielsen. Highly sparse representations from dictionaries are unique and independent of the sparseness measure. *Appl. Comput. Harm. Anal.*, 22(3):335–355, May 2007.

[40] J. Haupt and R. Nowak. Signal reconstruction from noisy random projections. *IEEE Trans. Information Theory*, pages 4036–4048, September 2006.

[41] P. Indyk and M. Ruzic. Fast and effective sparse recovery using sparse random matrices. 2008. avialable on arxiv.

[42] S. Jafarpour, W. Xu, B. Hassibi, and R. Calderbank. Efficient compressed sensing using high-quality expander graphs. available online at http://www.dsp.ece.rice.edu/cs/.

[43] G. James, P. Radchenko, and J. Lv. Dasso. Dasso: Connections between the dantzig selector and lasso. *J. Roy. Statist. Soc. Ser. B*, 71:127142, 2009.

[44] V. Koltchinskii. The dantzig selector and sparsity oracle inequalities. *Bernoulli*, 15(3):799–828, 2009.

[45] J. Mairal, F. Bach, J. Ponce, Guillermo Sapiro, and A. Zisserman. Discriminative learned dictionaries for local image analysis. *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[46] I. Maravic and M. Vetterli. Sampling and reconstruction of signals with finite rate of innovation in the presence of noise. *IEEE Trans. on Signal Processing*, 53(8):2788–2805, August 2005.

[47] N. Meinshausen, G. Rocha, and B. Yu. Discussion: A tale of three cousins: Lasso, l2boosting and dantzig. *Ann. Statist.*, 35(6):2373–2384, 2007.

[48] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246270, 2009.

[49] O. Milenkovic, R. Baraniuk, and T. Simunic-Rosing. Compressed sensing meets bionformatics: a new DNA microarray architecture. *Information Theory and Applications Workshop*, 2007.

[50] D. Needell and J. A. Tropp. CoSaMP: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

[51] D. Needell and R. Vershynin. Unifrom uncertainly principles and signal recovery via regularized orthogonal matching pursuit. *Foundations of Computational Mathematics*, 9(3):317–334, 2009.

[52] F. Parvaresh and B. Hassibi. Explicit measurements with almost optimal thresholds for compressed sensing. *IEEE ICASSP*, Mar-Apr 2008.

[53] F. Parvaresh, H. Vikalo, S. Misra, and B. Hassibi. Recovering sparse signals using sparse measurement matrices in compressed dna microarrays. *IEEE Journal of Selected Topics in Signal Processing*, 2(3):275–285, June 2008.

[54] G. Pisier. Probabilistic methods in the geometry of banach spaces. *Springer Lecture Notes*, 1206, 1986.

[55] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solution of linear matrix equations via nuclear norm minimization. 2007. available online at http://www.dsp.ece.rice.edu/cs/.

[56] F. Rodriguez and G. Sapiro. Sparse representations for image classification: Learning discriminative and reconstructive non-parametric dictionaries. 2008. available online at http://www.dsp.ece.rice.edu/cs/.

[57] J. Romberg. Imaging via compressive sampling. *IEEE Signal Processing Magazine*, 25(2):14–20, 2008.

[58] M. Rudelson and R. Vershynin. Geometric approach to error correcting codes and reconstruction of signals. *International Mathematical Research Notices*, 64:4019 – 4041, 2005.

[59] R. Saab, R. Chartrand, and O. Yilmaz. Stable sparse approximation via nonconvex optimization. *ICASSP, IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Apr. 2008.

[60] V. Saligrama and M. Zhao. Thresholded basis pursuit: Quantizing linear programming solutions for optimal support recovery and approximation in compressed sensing. 2008. available on arxiv.

[61] M. Stojnic. A rigorous geometry-probability equivalence in characterization of ℓ_1-optimization. available at arXiv.

[62] M. Stojnic. Upper-bounding ℓ_1-optimization weak thresholds. available at arXiv.

[63] M. Stojnic. Various thresholds for ℓ_1-optimization in compressed sensing. *submitted to IEEE Trans. on Information Theory*, 2009. available at arXiv:0907.3666.

[64] R. Tibshirani. Regression shrinkage and selection with the lasso. *J. Royal Statistic. Society*, B 58:267–288, 1996.

[65] J. Tropp. Just relax: Convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, March 2006.

[66] J. Tropp and A. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. on Information Theory*, 53(12):4655–4666, 2007.

[67] J. A. Tropp. Greed is good: algorithmic results for sparse approximations. *IEEE Trans. on Information Theory*, 50(10):2231–2242, 2004.

[68] S. van de Geer. High-dimensional generalized linear models and the lasso. *Ann. Statist.*, 36(2):614–645, 2008.

[69] H. Vikalo, F. Parvaresh, and B. Hassibi. On sparse recovery of compressed dna microarrays. *Asilomor conference*, November 2007.

[70] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. *Proc. Allerton Conference on Communication, Control, and Computing*, September 2006.

[71] J. Wright and Y. Ma. Dense error correction via ell-1 minimization. available online at http://www.dsp.ece.rice.edu/cs/.

[72] W. Xu and B. Hassibi. Efficient compressive sensing with determinstic guarantees using expander graphs. *IEEE Information Theory Workshop*, September 2007.